Faculty of Biological and Environmental Sciences
Doctoral Program in Integrative Life Sciences (ILS)
University of Helsinki
Helsinki, Finland

BAYESIAN ADVENTURES AMONG

HUMAN MITOCHONDRIAL LINEAGES

Sanni Översti

DOCTORAL DISSERTATION

To be presented for public discussion with the permission of the Faculty of Biological and Environmental Sciences of the University of Helsinki on the 26th of April, 2021 at 12 o'clock. The defence is open for audience through remote access.

Helsinki 2021

ISBN 978-951-51-7170-2 (Print)
ISBN 978-951-51-7171-9 (PDF)

Unigrafia
Helsinki 2021

Cover designed by M. Al-Soub and S. Översti.

The Faculty of Environmental and Biological Sciences uses the Urkund system (plagiarism recognition) to examine all doctoral dissertations.

Supervisors

Professor Päivi Onkamo
Department of Biology, University of Turku, Finland

and

Docent Jukka Palo
Department of Forensic Medicine, University of Helsinki, Finland

Reviewers

Professor Agnar Helgason
Department of Anthropology, University of Iceland, Iceland

and

Dr Boris A. Malyarchuk
Institute of Biological Problems of the North, Russian Academy of Sciences, Russia

Opponent

Professor Guido Barbujani
Department of Life Sciences and Biotechnology, Ferrara University, Italy

Members of the thesis advisory committee

Professor Niklas Wahlberg
Department of Biology, Lund University, Sweden

and

Professor Jukka Corander
Department of Biostatistics, University of Oslo, Norway

TIIVISTELMÄ

Mitokondrio on tuman ulkopuolella sijaitseva soluelin, jonka pääasiallisena tehtävänä on vastata solujen energia-aineenvaihdunnasta. Mitokondrioilla on muusta perimästä erillinen rengasmainen genominsa, mitokondriaalinen DNA (mtDNA), joka periytyy äidiltä kaikille jälkeläisille. Vaikka mtDNA muodostaakin vain pienen osan yksilön koko perimästä, sitä on hyödynnetty populaatiogeneettisissä tutkimuksissa laajalti: jäljittämällä äitilinjoja ajassa taaksepäin pääsemme tarkastelemaan naisten väestöhistoriaa.

Yksilön mtDNA-sekvenssiä kutsutaan hänen mitokondriaaliseksi haplotyypikseen. Haplotyypit jaotellaan muuntelun samankaltaisuuden perusteella haploryhmiin, joiden maantieteellinen jakautuneisuus nykyväestöissä tunnetaan hyvin. Haploryhmien yleisyydet eri väestöissä kertovat ihmisten liikkeistä ja väestösisäisen muuntelun määrän perusteella on mahdollista tehdä päätelmiä myös populaation koon vaihteluista. Viime vuosikymmeninä muinais-DNA-tutkimus on valottanut lukuisten mitokondriolinjojen alkuperää ja tiedämme, että esimerkiksi Euroopassa huomattava enemmistö kivikautisista metsästäjä-keräilijöistä edusti haploryhmää U. Maatalouden levitessä Lähi-Idästä noin 10,000 vuotta sitten saapui Eurooppaan kulttuurisen innovaation mukana lukuisia uusia haploryhmiä, kuten linjat H, J, K sekä T. Yhä tänä päivänä eurooppalaiset edustavat näitä eri alkuperää olevia haploryhmiä suomalaisten mitokondriaalisen rakenteen muistuttaessa pitkälti muita Euroopan väestöjä. Muusta genomista poiketen mitokondrioissamme ei ole nähtävissä eurooppalaisittain merkittävää itäistä vaikutusta. Niin ikään muussa perimässä selkeästi erottuvaa Suomen sisäistä geneettistä itä-länsi -jakoa ei ole toistaiseksi havaittu äitilinjoissamme.

Tämä väitöskirja tarkastelee nykysuomalaisten mitokondriaalista koostumusta Bayesilaisessa viitekehyksessä. Tutkimuksissa havaittiin, että vaikka päähaploryhmiemme jakauma onkin hyvin eurooppalainen, alahaploryhmistämme jopa kolmasosaa ei ole toistaiseksi löydetty Suomen ulkopuolelta lainkaan tai vain hyvin harvoin. Osittamalla äitilinjamme tähän ‘suomalaiskomponenttiin’ sekä muihin haploryhmiin, näemme toisistaan poikkeavat väestöhistoriat. Suomalaiskomponentin perusteella populaatiokoko on ollut pitkään pieni ja kasvanut merkittävästi vasta muutaman sata vuotta, kun taas muut mtDNA-linjat tuottavat hyvin ”eurooppalaisen” tuloksen: väestö on alkanut kasvaa merkittävästi tuhansia vuosia sitten. Muuntelun kerroksellisuuden lisäksi mtDNA-linjoissamme havaittiin maan sisäistä variaatiota: muinaisille metsästäjä-keräilijöille ominaiset linjat (U ja H) olivat Itä- ja Pohjois-Suomessa muuta maata yleisempiä kun taas maanviljelijöihin liitetyt haploryhmät (H, J, K, T) olivat yleisimmillään maan etelä- ja länsiosissa. Havainnot mtDNA-linjojen Suomen sisäisestä rakenteesta sekä muista eurooppalaisista erottuvasta

‘suomalaiskomponentista’ ovat linjassa sen kanssa, mitä suomalaisten taustoista tiedetään aiempien, esimerkiksi Y-kromosomianalyyseihin ja arkeologisiin löytöihin perustuvien, tutkimusten perusteella.

Väitöskirja käsittelee lisäksi evolutiivisten muuntelunopeuksien vaihtelua mitokondrion alahaploryhmien U2, U4, U5a sekä U5b välillä. Pääosin läntisen Euraasian alueelta koottujen muinaisten sekä modernien mtDNA-sekvenssien perusteella haploryhmän U5b substituutionopeus osoittautui merkittävästi muita tarkasteltavia linjoja alhaisemmaksi. Selitys havaittuun vaihteluun löytynee, ainakin osittain, linjojen läpikäymistä erilaisista populaatioprosesseista: haploryhmät U4 ja U5a on yhdistetty Pronssikaudella Euroopassa tapahtuneeseen nopeaan ja voimakkaaseen väestön leviämiseen, kun taas linjan U5b vallitsevuus vaikuttaa pysyneen maltillisen alhaisena vuosituhansien ajan. Eroavaisuudet substituutionopeuksissa vaikuttavat myös alahaploryhmien ikäarvioihin: aiemmista tutkimuksista poiketen linja U5b vaikuttaisi eriytyneen yhteisestä kantamuodosta huomattavasti linjaa U5a aiemmin. Muuntelunopeuksien eroilla on siis merkittävä roolinsa myös linjojen erkaantumisten sekä erilaisten demografisten tapahtumien ajoittamisessa. Vastaavaa evolutiivisten nopeuksien vertailua ei tiettävästi ole aiemmin toteutettu ihmisen mtDNA-linjoille, mutta tutkimuksen tulokset osoittavat haploryhmäkohtaisten substituutionopeuksien tarkastelun tärkeyden.

ABSTRACT

A mitochondrion is a cytoplasmic organelle responsible for the energy production of the eukaryotic cells. Mitochondria contain their own genome, mitochondrial DNA (mtDNA), which is a double-stranded circular molecule. Due to mitochondria's essential role in metabolism, a cell can contain hundreds of thousands of copies of mitochondrial DNA, depending upon the cell's energy requirements. In mammals, mtDNA is generally maternally inherited, meaning that it is transmitted from a mother to all of her descendants. Although mtDNA constitutes only a small fraction of the cell genome, it has several qualities which make it widely used in population genetic studies such as uniparental inheritance, and the fact that the mitochondrial genome does not recombine. Moreover, mtDNA has a mutation rate ten times higher than that of the nuclear genome and therefore allows us to trace back matrilineal lineages through generations and subsequently make inferences about maternal ancestors.

The human mtDNA sequence consists of approximately 16,570 base pairs and also contains both a coding region and a non-coding control region, the latter constituting around 7% of the whole mtDNA genome. Since mtDNA does not undergo recombination, an individual's mitochondrial haplotype can be determined simply by the direct sequencing of target amplicons. Haplotypes containing certain defining variants are considered to be descendants of a common ancestor and are classified into haplogroups. The geographical distribution of haplogroups among contemporary populations is well-known – for instance, the majority of Europeans exhibit mitochondrial lineages H, U, J, K, T and V. Ancient DNA research has uncovered that lineage U was already highly prevalent among the earliest hunter-gatherer settlers of Europe, whereas the gradual spread of agriculture from the Near East that started approximately 10,000 years ago brought along new people and hence also novel mitochondrial lineages (H, J, K and T). Previous studies have stated that compared with other European populations, contemporary Finns do not seem to be an exception in terms of mitochondrial genome pool. This is rather surprising, since other genetic markers have revealed that contemporary Finns are characterized by a strong Eastern genetic influence and are distinguishable from other Europeans. Moreover, the evident East-West distinction within Finland, apparent in Y-chromosomes and autosomes, has not been previously identified in the mtDNA.

This thesis outlines the mitochondrial DNA variation among present-day Finns in a Bayesian framework. The aims were to evaluate if Finns display a homogeneous geographical distribution of haplogroups and if the mtDNA composition of Finns resembles that of other European populations, as previously suggested. While no spatial differences have previously been detected in the mitochondrial haplogroup frequencies within Finland, a clear

geographical distinction arose when clustering haplogroups into 'hunter-gatherer' (U and V) and 'farmer' associated lineages (H, J, K and T). Whereas the farmer related haplogroups were notably more common in Southwestern Finland, the hunter-gatherer lineages had higher densities in the Northeastern parts compared to the Southwest. Furthermore, utilization of the complete mitochondrial genomes allowed for reassessing the Finnish mtDNA pool on a larger scale. One third of the subhaplogroups in Finland today were characteristic only of Finns, i.e. these lineages were virtually absent from other populations. When further partitioning the Finnish samples based on their inclusion in 'local' and 'non-local' lineages, two notably different demographic trajectories were obtained. The population history for Finn-characteristic lineages was more in accordance with what is known through other data types, such as Y-chromosomal and archaeological data. In general, the observed geographical within-country deviation in the Finnish mtDNA pool and the high proportion of Finn-characteristic lineages reflected the signals reported from other genetic markers.

Alongside Finnish mtDNA, this thesis explores the molecular rate variation among the different subhaplogroups of lineage U. Unexpectedly, a noteworthy discrepancy emerged from the tip calibrated phylogenies: haplogroup U5b had a notably lower substitution rate when compared to U2, U4 and U5a. This lineage-specificity in the rates most likely arose, at least to some extent, from differences in past population dynamics. In particular, U4 and U5a have been associated with the rapid population expansion which occurred during the Bronze Age, whereas the frequency of U5b has remained rather stable. Subsequently, the observed rate of deviation influences the divergence estimates for subhaplogroups, suggesting that U5b emerged considerably earlier than U5a. Since molecular rates are fundamental to several population genetic analyses and the timing of divergence and demographic events relies heavily on the rate used, more attention should be paid to the interlineage molecular rate variation.

The results of this thesis demonstrate not only the importance of using complete mtDNA genomes and the appropriate molecular rate, but also the relevance of approaching the data from new angles when assessing the demographic past of mitochondrial lineages.

CONTENTS

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following original publications (I-II) and manuscript (III) which are referred to in the text by their Roman numerals

I. Neuvonen Anu M., Putkonen Mikko, Översti Sanni, Sundell Tarja, Onkamo Päivi, Sajantila Antti, Palo Jukka U. Vestiges of an Ancient Border in the Contemporary Genetic Diversity of North-Eastern Europe. *PLoS ONE*, 10(7):e0130331 (2015)

II. Översti Sanni, Onkamo Päivi, Stoljarova Monika, Budowle Bruce, Sajantila Antti, Palo Jukka U. Identification and Analysis of mtDNA Genomes Attributed to Finns Reveal Long-stagnant Demographic Trends Obscured in the Total Diversity. *Scientific Reports*, 7:6193 (2017)

III. Översti Sanni, Palo Jukka U. Variation in the substitution rates among the human mitochondrial haplogroup U sublineages. *Manuscript*.

Publication I has earlier appeared as part of the doctoral thesis of Anu M. Neuvonen (2017, University of Helsinki).

AUTHOR CONTRIBUTIONS

	Study I	**Study II**	**Study III**
Conceived and designed the study	AN MP SÖ TS PO AS JP	SÖ, JP, PO	SÖ
Collected the data	AN MP SÖ TS PO AS JP	SÖ, JP	SÖ
Provided the data	AN MP SÖ TS PO AS JP	MS, BB	-
Performed the data analysis	AN MP SÖ TS PO AS JP	SÖ, JP, PO	SÖ
Wrote the manuscript	AN MP SÖ TS PO AS JP	SÖ, JP	SÖ, JP
Supervised the study	JP, AS	JP, PO, AS	JP

AN = Anu Neuvonen
AS = Antti Sajantila
BB = Bruce Budowle
JP = Jukka Palo
MP = Mikko Putkonen
MS = Monika Stoljarova
PO = Päivi Onkamo
SÖ = Sanni Översti
TS = Tarja Sundell

ABBREVIATIONS

AD	Anno Domini (Common Era)
aDNA	Ancient DNA
BF	Bayes factor
bp	Base pair
BSP	Bayesian Skyline Plot
calYBP	Calibrated years before present
CRS	Cambridge Reference Sequence
CWC	Corded Ware culture
DNA	Deoxyribonucleic acid
DYS	DNA, Y-chromosome, unique Segment
EHG	Eastern hunter-gatherer
ESS	Effective sample size
FARM	Farmer associated mitochondrial haplogroups
FDH	Finnish disease heritage
GTR	General Time Reversible model
HKY	Hasegawa, Kishino, Yano mutation model
HPD	Highest posterior density
HUNT	Hunter-gatherer associated mitochondrial haplogroups
HVR1	Hypervariable region 1
HVR2	Hypervariable region 2
JC69	Jukes Cantor mutation model
kya	Thousands of years ago
LGM	Last Glacial Maximum
MCMC	Markov chain Monte Carlo
ML	Maximum likelihood (methods)
MRCA	Most Recent Common Ancestor
mRNA	Messenger ribonucleic acid
MSA	Multiple sequence alignment
mtDNA	Mitochondrial DNA
N_e	Effective population size
NE	Northeastern (Finland)
NRY	Non-recombining part of Y
OTU	Operational taxonomic unit
PAR	Pseudoautosomal region
pInv	Proportion of invariant sites
rCRS	Revised Cambridge Reference Sequence
rRNA	Ribosomal ribonucleic acid
RNA	Ribonucleic acid
STR	Short tandem repeat
SW	Southwestern (Finland)
TCW	Typical Combed Ware culture
TN93	Tamura Nei mutation model
WHG	Western hunter-gatherer
ybp	Years before present
Y-SNP	Y-chromosomal single nucleotide polymorphism
Y-STR	Y-chromosomal short tandem repeat
^{14}C	Radiocarbon

REVIEW OF THE LITERATURE

1 INTRODUCTION TO GENETICS

For the majority of organisms, including humans, genetic information is encoded in deoxyribonucleic acid (DNA) which consists of four nucleotide bases: adenosine, cytosine, guanine and thymine (A, C, G and T, respectively). Nucleotides attach together to form a polymeric DNA strand and subsequently two complementary DNA strands can join together, establishing a double-stranded helical structure. The double-stranded DNA molecule is densely coiled and forms tightly packed chromosomes together with assisting proteins. Humans have 23 chromosome pairs – one copy of each chromosome inherited from one parent – which makes us diploid organisms. The chromosomes are located in the nucleus of the cell and they consist of autosomes (chromosome pairs 1–22) and sex chromosomes, the latter determining the genetic sex of an individual. Humans have two types of sex chromosomes, X and Y, with females being XX and males XY. Since the Y-chromosome determines the genetic sex of males, it is paternally inherited, i.e. always passed on from father to son. In addition to nuclear chromosomes, a lesser amount of genetic information is also present in the mitochondria, known as mitochondrial DNA (mtDNA). This cytoplasmic organelle is responsible for the energy metabolism of the cell and it is maternally inherited, that is transmitted from a mother to her children (**Figure 1.** 'Modes of inheritance').

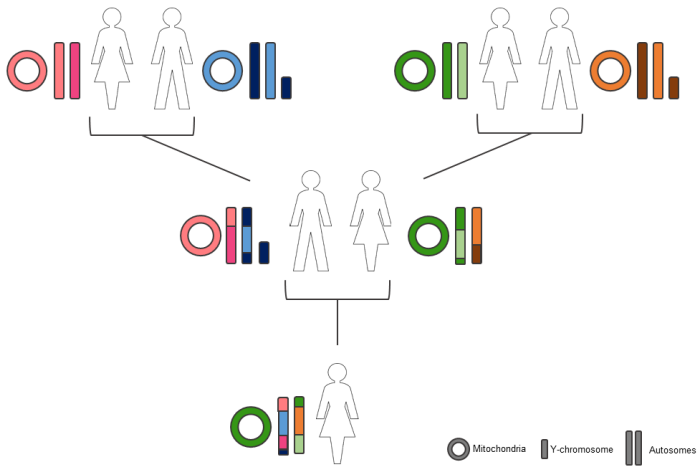

Figure 1 Modes of inheritance. Schematic illustration of a human pedigree with three generations. The majority of the genome is inherited from both parents through autosomal chromosomes, which undergo recombination. The Y-chromosome is always transmitted from father to son whereas a mother passes her mitochondrial DNA (mtDNA) to all of her children.

The haploid human genome is composed of approximately three billion base pairs (bp) and contains both functional and non-functional regions. The proportion of functional elements in the genome remains highly debated, with estimates varying from 8 to 80 % (Pennisi 2012; Rands *et al.* 2014; Graur 2017). Nevertheless, it has been estimated that the human genome constitutes approximately 20,000 protein-coding genes, covering only ~1% of the whole human genome (Abdellah *et al.* 2004; Piovesan *et al.* 2019). Out of these genes, only 37 reside in the mitochondria with the rest located in autosomal and sex chromosomes. Compared to most other organisms, the genetic variability within a human population is relatively low, which has been interpreted to be a signal of the comparatively young age of anatomically modern humans as a species and long-term small population size (Li and Sadler 1991; Jorde *et al.* 1998; Kaessmann *et al.* 1999; Osada 2015).

1.1 Processes shaping genetic diversity

A population is defined as a group of individuals belonging to the same species that occupy a defined geographical area and have the capability to breed. The genetic information carried by the individuals belonging to the population forms the population's gene pool. Within the population's gene pool, four processes can drive evolution, i.e. changes in the allele frequencies: mutation, migration, drift and selection (**Figure 2.** 'Processes shaping genetic diversity'). Of these evolutionary forces, mutation alone can create new alleles. In addition to mutation, recombination increases the genetic diversity of the population, as it creates new haplotype combinations, but it does not alter allele frequencies. However, since this thesis focuses on uniparental markers, which for the most part do not recombine, it will not be discussed here in more detail. For a general overview of processes shaping genetic diversity, see Jobling *et al.* 2014 pp. 132–165 and Klug *et al.* 2016 pp. 681–702.

1.1.1 Mutation

A mutation is considered to be any change in a nucleotide sequence and can range from single nucleotide changes to complete chromosomal rearrangements. Further, mutations can create new alleles, which are alternative forms of a gene or genetic loci at a specific chromosomal location. Various distinct mechanisms are responsible for mutations, such as errors during DNA replication or cell division. If a mutation occurs in the germline, it can be transmitted to subsequent generations and thus might contribute to a population's evolution. Depending on the mutation type and location in the genome, a mutation can be either beneficial, neutral, deleterious, or lethal. If a mutation occurs in a DNA sequence encoding for a gene, it might subsequently change the amino acid and thus also alter the protein sequence.

In contrast to these nonsynonymous mutations, due the redundancy of genetic code, the amino acid might remain unchanged regardless of nucleotide substitution. In most cases, these synonymous mutations occur in the last position of a three-nucleotide codon. Since the nonsynonymous mutations alter the protein sequence, they are frequently subjected to natural selection, whereas synonymous mutations are generally considered to be evolutionarily neutral.

1.1.2 Genetic drift

The random fluctuation in the allele frequencies between successive generations is known as genetic drift. This sampling error occurs in every finite natural population, but its influence is strongest in small populations. Eventually, it might result in the fixation of some alleles or complete loss of others, both resulting in a loss of allelic diversity of a population. There are two extreme cases of genetic drift: founder effect and population bottleneck. Founder effect refers to an event where a subset of a population colonizes a previously uninhabited area. As the gene pool of these migrants represents only a fraction of the genetic diversity of the original population, the average heterozygosity might be substantially less than in the source population. Population bottleneck refers to a temporally rapid reduction in the population size, caused by, for example, environmental change or lethal disease. This reduction results in a decline in genetic heterozygosity and the severity of this decline is further dependent on the length of the bottleneck in generations as well as the number of breeding individuals in the population (i.e., effective population size, see section 1.5.2.1. 'Coalescent based models') (Nei *et al.* 1975).

1.1.3 Migration

Migration is the movement of an individual or group of individuals between two or more populations. Migration from one habitat to another can be either active, such as walking or flying, or passive, such as dispersing along wind or water currents. If interbreeding occurs between the migrant and recipient populations, new alleles might be introduced into the gene pool of the receiving population's next generation. Consequently, this gene flow increases the genetic diversity of the recipient population. In addition, migration might also affect the genetic diversity of the source population by reducing allele diversity and thus causing decrease in heterozygosity. Nevertheless, gene flow prevents differentiation of populations as it counteracts the influences introduced by genetic drift.

1.1.4 Selection

Selection refers to variation in reproductive success caused by diversity in genotypes and hence also in phenotypes; individuals with different phenotypes possess varying capabilities to survive and reproduce in different living environments. This capability is measured by fitness, which describes an individual's genetic contribution to future generations. Selection can be positive, where alleles enhancing fitness are favored and subsequently frequencies for these alleles increase in the population. Conversely, negative selection counteracts alleles that reduce fitness and eventually these alleles might be removed from the population. Thus, selection is a process that alters allele frequencies in a non-random manner.

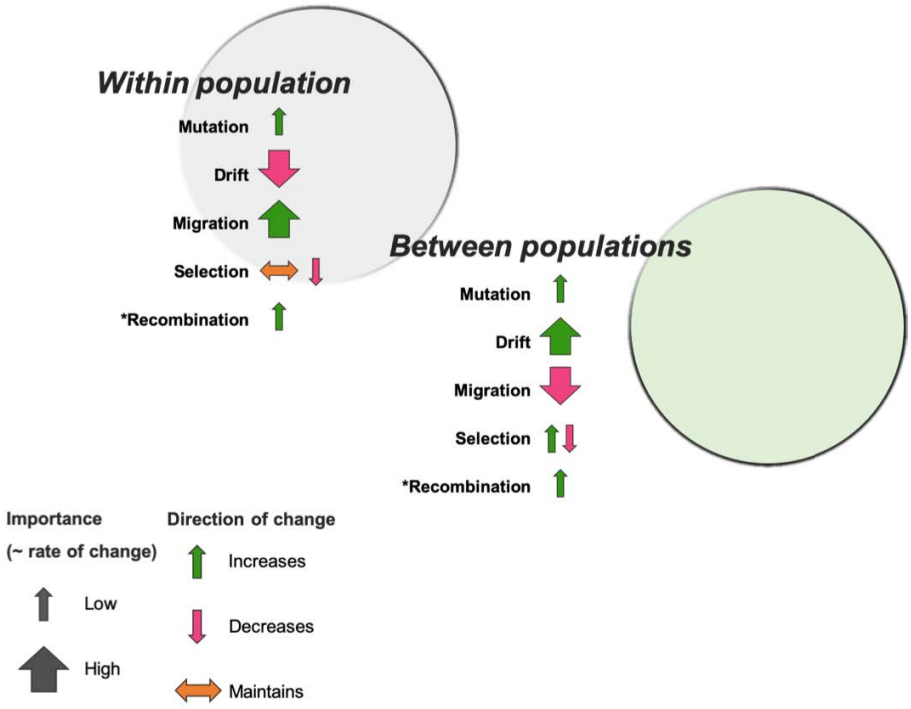

Figure 2 Processes shaping genetic diversity and the nature of their impact on genetic distances between populations and diversity within population. * Although recombination creates new combinations of alleles and therefore introduces more genetic variation into the population's gene pool, it doesn't change allele frequencies.

1.2 Uniparental markers in human population genetics

Since mitochondrial and Y-chromosomal DNA are inherited from only one parent to their descendants, these uniparental markers could be used to trace back matrilineal and patrilineal lineages respectively. As mitochondrial DNA does not recombine (Merriwether *et al.* 1991) and Y-chromosomal DNA is

largely non-recombining, mutations are the only source of variation detected in these markers. This makes it possible to obtain an individual's matrilineal or patrilineal lineage directly, and hence these markers can be used to infer sex-specific patterns in a population's past. However, as uniparental markers represent only two loci out of the whole genome, they cannot capture the whole evolutionary past of a population. Additionally, because of the recent rapid improvement in sequencing techniques and computational analysis methods, population genetics studies are shifting towards the usage of autosomal data. Nevertheless, uniparental markers are still routinely used both in population genetics and in forensic genetics due to their unique pattern of inheritance, well-established geographical distribution, and ease of genotyping.

1.2.1 Mitochondrial DNA

The mitochondrion is a cytoplasmic organelle with a suggested endosymbiotic bacterial origin, due to its genetic features more resembling bacterial genomes than the genomes in eukaryotic cells. These features include, among others, different genetic code when compared to the nuclear genome (Barrell *et al.* 1979) and lack of introns in the genes (Anderson *et al.* 1981). Since mitochondria are responsible for the energy production of the cell, their DNA is highly conserved among eukaryotic cells (Clayton 1992). Further, the number of mitochondrial genomes varies between cell types according to the energy demand of the cell. In mammals, the majority of cell types enclose up to 1,000 mitochondria, each containing approximately 2-10 mtDNA copies (Robin and Wong 1988). In mature oocytes, the number of mitochondria can encompass several hundreds of thousands copies of mtDNA (see Monnot *et al.* 2013 and references therein).

A generally accepted view is that in mammals, mtDNA is maternally inherited, since the mitochondria of the sperm are selectively degraded during egg fertilization (Sutovsky *et al.* 1999). However, for some species the purely maternal inheritance of mtDNA has been debated (Ladoukakis and Zouros 2001) and a recent study has shown traces of biparental inheritance also in humans (Luo *et al.* 2018). Nevertheless, several features make mtDNA convenient for population level genetic analyses, such as its haploid genome (Hutchison *et al.* 1974), lack of recombination (Merriwether *et al.* 1991), and high copy number per cell (Pikó and Matsumoto 1976). In addition, in the mammalian genomes the mitochondrial mutation rate is approximately ten times higher than in the nucleus (Brown *et al.* 1979), which makes it possible to also identify relatively recent divergence events.

Mitochondrial DNA is a circular, two-stranded molecule with heavy (H) and light (L) chains, which differ in their base composition. In humans, the mitochondrial DNA sequence is approximately 16,569 base pairs long and it is further divided into a short non-coding control region and a coding region (**Figure 3.** 'Mitochondrial DNA'). The coding region contains 37 genes, out of which 13 encode proteins associated with the oxidative phosphorylation

pathway, 22 encode messenger ribonucleic acids (mRNA) associated with mitochondrial protein synthesis, and the last two genes produce ribosomal RNAs (rRNA) (Anderson *et al.* 1981).

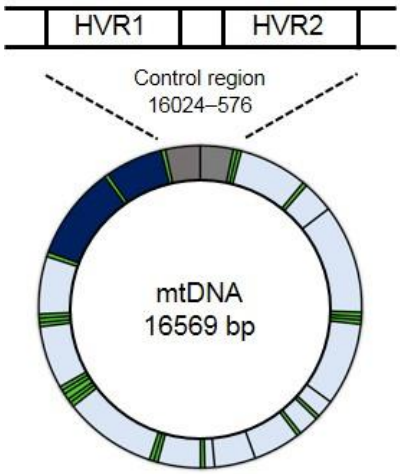

Figure 3 Mitochondrial DNA. The sequence is divided into a coding region (base pairs 577–16,023) and control region (base pairs 16,024–576, dark grey). The latter is subsequently divided into hypervariable regions 1 and 2 (HVR1 and HVR2). HVR1 consists of base pairs 16,024–16,385 and HVR2 base pairs 73–340. The coding region encodes 22 messenger-RNAs involved in protein synthesis (green), two ribosomal-RNAs (dark blue) and 13 metabolism related proteins (light blue).

1.2.2 Mitochondrial haplotypes and haplogroups

The first complete human mitochondrial genome sequenced (Anderson *et al.* 1981) was established as a reference sequence and is known as the Cambridge Reference Sequence (CRS). The numbering of positions along the mitochondrial genome is based on this 16,569 bp long sequence. Over the years, the CRS has been reanalyzed and 11 rare polymorphisms and sequencing errors have been corrected, yielding the revised Cambridge Reference Sequence rCRS (Andrews *et al.* 1999, GenBank sequence number NC_012920). Due to the haploid mode of inheritance of mtDNA and the lack of recombination, an individual's mitochondrial haplotype can be directly obtained from the sequence. Furthermore, individuals bearing the same polymorphism(s) are considered to be descendants of the same maternal ancestor. It is generally agreed that some rare polymorphisms define lineage divergence events in the human mitochondrial tree and based on these mutations, evolutionarily similar haplotypes are classified into so-called haplogroups, which are designated with letters from A to Z (**Figure 4.** 'Schematic illustration of mitochondrial haplogroup tree'). Moreover, each main haplogroup consists of subhaplogroups, marked with combinations of letters and numbers such as U5, U5b, U5b1, and so forth. The mitochondrial

haplogroup of an individual is determined by comparing its haplogroup-defining polymorphisms against the rCRS, which represents haplogroup H2a2.

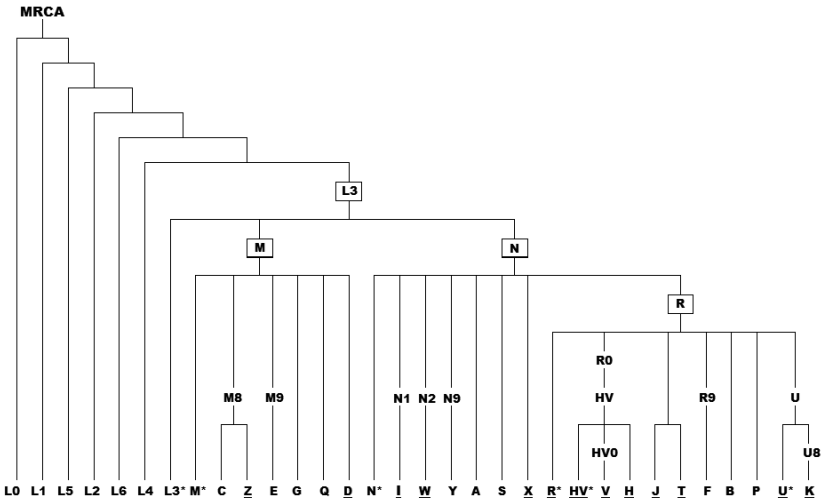

Figure 4 Schematic illustration of the mitochondrial haplogroup tree. Figure is modified according to the PhyloTree version 17 (https://www.phylotree.org/, May 2019) (van Oven and Kayser 2009). As also defined in PhyloTree, haplogroup symbols indicated with an asterisk represent all other subhaplogroups of a particular lineage, for instance U* includes U1, U2, U3, U4, U5, U6, U7 and U9. MRCA stands for the most recent common ancestor. Haplogroups that are mainly observed among Finns are underlined.

1.2.3 Geographical distribution of mitochondrial haplogroups

In modern human populations, geographical differences in haplogroup frequencies are detected primarily between continents. Africans, displaying the highest level of continental mtDNA variation (Chen *et al.* 1995), belong mainly to the haplogroups L0, L1, L2 and L3. For L0, the haplogroup-defining mutations distinguishing it from the other haplogroups are considered to be the oldest of all existing mtDNA variants (Soares *et al.* 2009, Behar *et al.* 2012). Populations outside of Africa belong predominantly to macrohaplogroups M and N, which diverged from African lineage L3. Macrohaplogroup M includes lineages C, D, E, G, Q and Z, which are distributed mainly in Southern and Eastern Asia and also on the American continent. Macrohaplogroup N consists of A, I, S, W, X, Y and R, of which R is further divided into haplogroups B, F, J, P, T, HV, H, V, U, and additionally in U's subhaplogroup K. The vast majority of Europeans belong to the haplogroups H, I, J, K, T, U, V, and X (Torroni *et al.* 1996), with haplogroups H and U being most prevalent in the majority of European populations (**Figure 5.** 'Mitochondrial haplogroup frequencies in Europe'). Whereas the

distribution of H is relatively uniform across Europe, U shows a geographical pattern being more common in Northeastern Europe. U reaches its highest frequency in present-day Saami of up to 48% (Sajantila *et al.* 1996; Tambets *et al.* 2004).

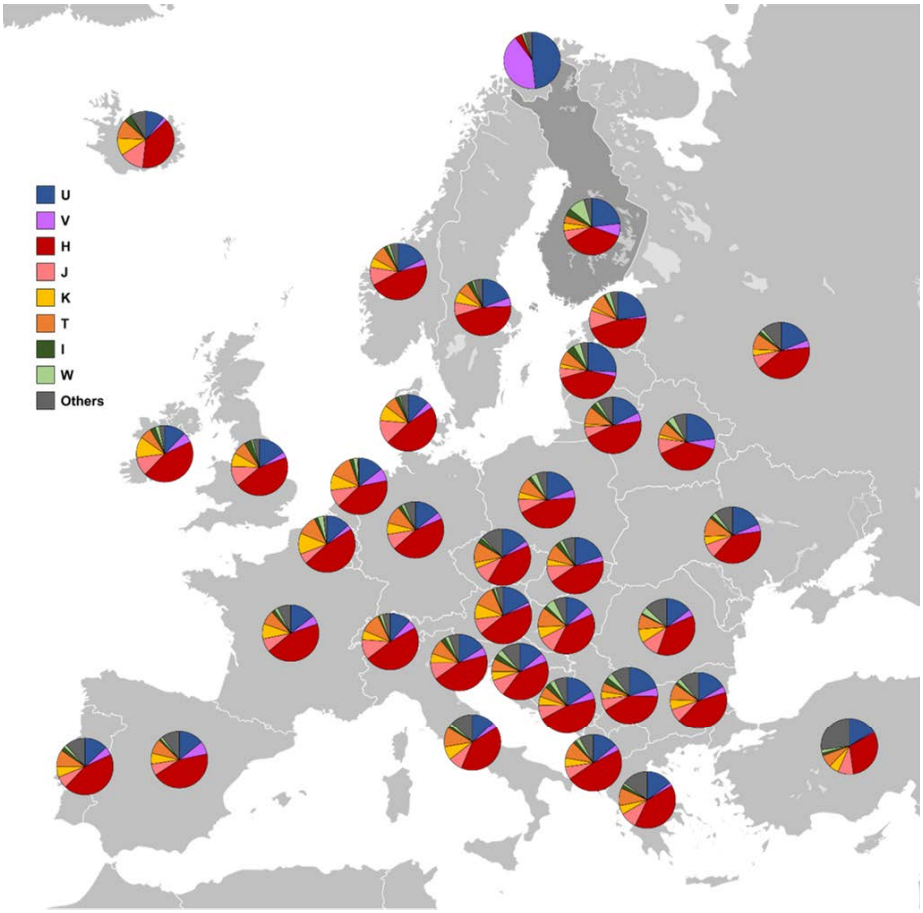

Figure 5 Mitochondrial haplogroup frequencies in Europe. Frequencies obtained from eupedia.com (May 2019). Finland is presented in darker grey. Similar to other European populations, Finns exhibit mainly haplogroups H, I, J, K, T, U, V, W, and X.

1.2.4 Y-chromosomal DNA

The size of the Y-chromosome in humans is approximately 60 million base pairs (Human Genome Assembly GRCh38, https://www.ncbi.nlm.nih.gov/grc/human/data?asm=GRCh38), which makes it one of the smallest chromosomes in the human genome. Additionally, the Y-chromosome is extremely gene-poor, containing only around 60 protein-coding genes, most of them involved in the determination of male specific traits. The majority of the human Y-chromosome consists of a non-

recombining region (NRY) flanked by pseudoautosomal regions (PAR), which can exchange genetic material with complementary regions in the X-chromosome. This capability suggests that X and Y evolved from an autosomal chromosomal pair, but Y has lost several ancestral gene functions during its evolution (for review see Graves 1995; Quintana-Murci and Fellous 2001). Approximately 30 megabases of the non-recombining region is composed of densely coiled repeat-rich heterochromatin, which makes it challenging to assemble around half of the Y-chromosomal sequence (for review see Bachtrog and Charlesworth 2001).

The non-recombining part of the Y-chromosome is directly transmitted along the paternal lineage and hence the only source of variation is mutation. Y-chromosomal mutations mainly consist of two types of polymorphisms: slowly mutating bi-allelic single nucleotide polymorphisms (Y-SNP) and fast evolving multi-allelic short tandem repeats (Y-STR), also known as microsatellites. STRs are tandem repeats of short DNA motifs, usually between 1–7 base pairs, with a high degree of polymorphism as the copy number of the motif typically varying somewhere between 8 and 30. This extensive variability in the number of repeats makes it possible to distinguish two paternally unrelated males from each other, which makes them essential for forensic research but also highly practical for population genetic studies. In forensic research, panels covering 9 to 27 standard Y-STR markers have been widely utilized but in addition, several hundred other Y-chromosomal tandem repeats have been identified (for overview see Kayser 2017). All Y-STR markers have a unique identification symbol, commonly designated with DYS numbers (abbreviation from: DNA, Y-chromosome, unique Segment).

The mechanism responsible for the variation in the microsatellite copy number is known as 'replication slippage', caused by the misalignment of the DNA-strands during the replication process. Upon a mutation, the number of motifs can increase or decrease by the repeat length of one (one-step mutation) or more (multi-step mutation) (see Sainudiin *et al.* 2004 and references therein). Studies have revealed that over 90% of the Y-STR mutations are single-step mutations and only a minor fraction include allele expansion or contraction by more than one repeat unit (Goedbloed *et al.* 2009; Claerhout *et al.* 2018; Boattini *et al.* 2019). In tri- and tetranucleotide motifs it is mostly single-step mutations which occur, whereas dinucleotide motifs are more prone to multi-step mutations (Willems *et al.* 2016). Further, repeat gains are slightly more common than losses (57% and 42%, respectively) (Goedbloed *et al.* 2009). It has also been shown that loci with longer allele length tend to mutate more often (Ellegren 2000) and that the rate of repeat number reduction increases exponentially with the allele length (Xu *et al.* 2000; Dupuy *et al.* 2004).

1.2.5 Y-chromosomal haplotypes and haplogroups

Since recombination does not occur in the NRY, the haplotype of an individual is directly obtained from his Y-chromosomal sequence. Further, Y-chromosomal haplogroups can be assigned by the bi-allelic haplogroup-defining mutations (Y-SNPs), by Y-STR alleles, or by combination of these. Y-chromosomal lineages are traditionally confirmed based on the single nucleotide polymorphisms, but since the mutation rate for Y-SNPs is considerably lower than for the Y-STRs (see table 1 in Balanovsky 2017), STR markers are routinely used to define the sublineages and to distinguish fine-resolution variation. Y-chromosomal haplogroups can be labeled either with the defining SNP or with a combination of letters and numbers, such as I-DF29 or I1a (**Figure 6.** 'Schematic illustration of Y-chromosomal haplogroup tree).

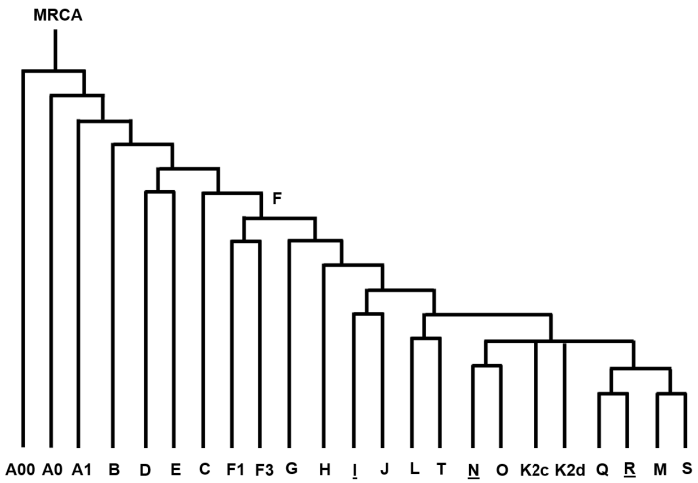

Figure 6 Schematic illustration of Y-chromosomal haplogroup tree. Figure is modified according to the YFull YTree v8.09.00 (www.yfull.com/tree/, October 2020). Haplogroups that are mainly observed among Finns are underlined.

1.2.6 Geographical distribution of Y-chromosomal haplogroups

Haplogroups A00, A0 and A1 in the root of the Y-chromosomal genealogical tree are predominantly present in Africa (http://phylotree.org/Y/tree/index.htm, Van Oven *et al.* 2014) (**Figure 6.** 'Schematic illustration of Y-chromosomal haplogroup tree). Similar to mitochondrial DNA, Y-STR diversity decreases with distance from Africa (Shi *et al.* 2010), strengthening the hypothesis of the African origin of modern humans. The tree further deviates into sub-clades B, C, D, E, and F, with different geographical distributions. Haplogroups B and E are mainly concentrated on the African continent, whereas lineage D is mainly prevalent in East Asia and C is widely distributed across Asia and the Pacific. Europeans belong mainly to the sublineages that diverged from the main haplogroup F,

such as I, J, N, R. and T (**Figure 7.** 'Y-chromosomal haplogroup frequencies in Europe').

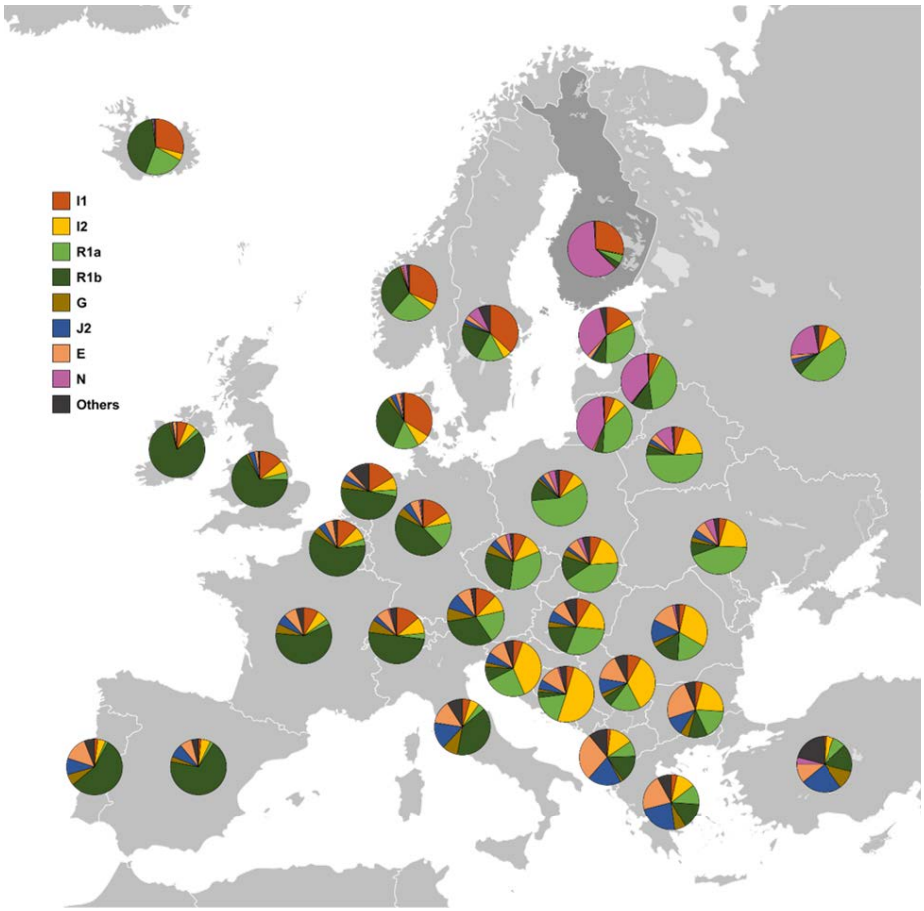

Figure 7 Y-chromosomal haplogroup frequencies in Europe. Frequencies obtained from eupedia.com (May 2019). Finland is presented in darker grey. Finns belong mainly to the haplogroups N (N-M231), I1 (I-DF29), R1a (R-M420) and R1b (R-M343).

1.3 Phylogenetic trees

Phylogenies, commonly visualized as phylogenetic trees, are used to study evolutionary relationships among taxa (i.e., operational taxonomic units, OTUs). Phylogenies can also be used to address questions concerning the migration patterns of taxa or demographic changes within a population. The two main classes of phylogenetic trees are the unrooted tree, where only the relatedness of the taxa is described, and the rooted tree, which additionally characterizes the direction of evolution. This means that the latter also makes assumptions about the common ancestries and hence the root of the tree can

be interpreted as the most recent common ancestor (MRCA) for all taxa included in the tree. Despite phylogenetic trees being widely used to describe the evolutionary history of a set of taxa, it must be noted that evolution doesn't always proceed in a bifurcating, tree-like manner. Events such as horizontal gene transfer and recombination violate the assumption of binary-branching evolution and thus phylogenetic networks are considered to provide a more sophisticated characterization of taxa relatedness (for a review of phylogenetic networks in evolutionary studies see Huson and Bryant 2006).

Methodologies for the reconstruction of phylogenetic trees are based on either genetic distances, such as the neighbor-joining method (Saitou and Nei 1987), or on characters, such as the methods relying on Bayesian inference. Concerning the distance-based methods, the pairwise sequence distances are calculated according to a pre-determined substitution model (see section 1.5.1 'Priors on the DNA substitution model') and the resulting matrix is used to reconstruct the phylogenetic tree with a chosen algorithm. In turn, the distance-based neighbor-joining method utilizes a cluster algorithm to produce the phylogenetic tree under the principle of minimum evolution (Saitou and Nei 1987). Tree construction begins with a star-like tree, for which the OTUs are clustered based on the distance matrix, starting from the two most closely related taxa. The combination of these two taxa is then considered as a single OTU and the distance matrix is recalculated. The same steps are repeated until the tree topology is fully resolved.

By contrast, the character-based methods rely on an optimality criterion and use the multiple sequence alignment (MSA) directly. In short, the MSA is constructed for three or more molecular sequences that are considered to be descendants of a single ancestral sequence and the outcome serves as a starting point e.g. for phylogenetic analyses. The MSA determines similarities and dissimilarities among taxa by identifying homologous regions between query sequences. Assuming that the observed differences, such as mutations, insertions and deletions, are acquired since the divergence from the ancestral sequence, it is possible to estimate the evolutionary distances between the taxa. Since aligning multiple sequences might be a computationally intensive process, several heuristic methods such as MUSCLE and MAFFT (Edgar 2004; Katoh and Standley 2013, respectively) have been developed to provide an approximate solution for building up an MSA. Despite the availability of a variety of different MSA algorithms and even the existence of tools for evaluating the quality of the alignment, recognition of poorly aligned regions might be challenging. However, as the accuracy of the alignment has a huge impact on the correctness and interpretation of the phylogeny (Ogden & Rosenberg 2006), subjective and experience-based inspection of the alignment is highly recommended. For a review of different MSA methods, see for example Chowdhury and Garai 2017.

In character-based methods, all the sequences in the MSA are analyzed at the same time and each sequence position in the alignment is independently considered. Based on the similarities of the sites along the MSA, the method

then calculates a 'score' for the phylogenetic tree. Different character-based methods define this score differently; in the maximum parsimony method, it is the number of changes in the tree. All the possible phylogenetic trees are constructed based on the MSA and for each tree the number of character-state changes are calculated. The tree which produces the least number of changes ('minimum evolution') is then considered to explain the data the best. In maximum likelihood (ML) methods, the score is determined by likelihood, which is simply the probability of the observed sequences assuming a certain evolutionary model, such as a substitution model. The analysis produces phylogenetic trees with different topologies and hence also with distinct likelihoods. The tree topology with the highest likelihood is then considered to be the best description of the phylogenetic relationships of the sequences.

Whereas the ML methods aim to optimize the probability of the data given a model, in Bayesian inference the parameters of the model are random variables with pre-determined prior distributions (see section 1.4 'Bayesian methods in phylogenetics'). These prior distributions are then used together with the data to calculate the posterior distributions for each parameter included in the model and hence also for the trees. This thesis focuses on Bayesian methods in phylogenetics; for a description of other methods in molecular phylogenetics see the review by Yang and Rannala 2012.

1.4 Bayesian methods in phylogenetics

The Bayesian framework, which was introduced to phylogenetics in 1996 (Rannala and Yang 1996), is based on Bayes' theorem. Assuming that D is the data and θ is a set of parameters, Bayes' theorem takes the following form:

$$(1) \qquad P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)},$$

where the $P(\theta|D)$ is the probability distribution of the parameters given the data, $P(D|\theta)$ is the probability of the data given the model, $P(\theta)$ denotes the prior probability, and the denominator $P(D)$ is the probability of the data. In the phylogenetic framework, the data (D) is typically a multiple sequence alignment and θ is the collection of the parameters in the model, such as the parameters of the substitution model.

The $P(D|\theta)$, also known as the likelihood, is the probability of the observed data given the predetermined model. This model can, for instance, include substitution and tree models. Put simply, the likelihood is the probability that the observed data would have been obtained, assuming that a particular evolutionary model holds. The prior probabilities, $P(\theta)$, are distributions which describe a researcher's prior beliefs about the parameters in the model,

independent of the data. This prior probability could be, for example, a known estimate for a substitution rate. Setting these prior distributions should be done with caution since they might have a considerable effect on subsequent analyses and results (for a review, see Bromham *et al.* 2018). The marginal likelihood, *P(D)*, which describes the probability of the data, is typically challenging to determine. However, due to its nature of being constant during the analysis, it can be excluded from the calculations. Therefore, the posterior probability $P(\theta|D)$ can be considered as a combination of the prior beliefs and the information obtained from the data (likelihood).

Principally, the analysis starts by assuming the prior probabilities of the parameters included in the model and these prior beliefs are then updated according to the signals obtained from the data. The updated prior beliefs are the resulting posterior distributions for each parameter. Since the direct calculation of the posterior probabilities for the trees is generally computationally unachievable, Markov Chain Monte Carlo (MCMC) algorithms are utilized. MCMC is a simulation algorithm, which draws a large number of samples from each parameter's, posterior distribution. MCMC randomly chooses the starting tree and the starting values for parameters and then estimates the posterior probability for the tree, i.e. the likelihood weighted with the prior probabilities. Subsequently, the algorithm makes changes to one or more parameters and calculates the posterior probability of the new proposed tree. The algorithm then compares the ratio of these two posterior probabilities and if the new state has a higher probability than the preceding state, the new state is accepted and MCMC makes a new proposal for changes to one or more parameters. If the posterior probability of the new state is notably less than the current one, the algorithm stays in the current state and makes a new proposal. However, if the posterior probability is only slightly less than the current $P(\theta|D)$, the algorithm decides whether it should accept or reject the new state by drawing a random number from a uniform [0, 1] distribution. If this random number is smaller than the ratio calculated for the posterior probabilities of the current and proposed state, then the new state is accepted; otherwise the new state is rejected and MCMC continues making proposals. Usually, MCMC continues until a predetermined number of states is reached. For instance, in the analysis included in this thesis the number of states is set to 15,000,000-60,000,000 steps depending on the sample size. MCMC proceeds in the space of parameters as described and due to the nature of algorithm's state acceptance/rejection, each tree is sampled according to its posterior probability. For accepted states, the values of the parameters are logged and these values form the individual posterior distribution for each parameter included in the model. The most commonly used MCMC method in Bayesian phylogenetics is the Metropolis–Hastings algorithm (Metropolis *et al.* 1953; Hastings 1970).

The Bayesian phylogenetic framework has become one of the most popular inferences in phylogenetics as it allows complex evolutionary models with multiple parameters. Furthermore, Bayesian inference could be used not only

to estimate genealogies but also to evaluate the demographic and phylogeographic past of the population in a temporal framework, which is currently not possible in the traditional frequentist framework. Moreover, increases in computational power during the past few decades have made it possible to analyze large datasets and, in addition, implement more complex evolutionary models.

1.5 Priors for Bayesian inference

As described in the previous section, Bayesian methods require the specification of prior distributions which reflect the researcher's prior assumptions of the parameters included in the models used. Defining these prior distributions allows us to include any information we have on the underlying processes, independent of the data. However, implementation of improper priors might cause a bias in the posterior distributions of the model parameters and in the trees inferred. Since selection of proper priors is crucial to the Bayesian phylogenetic analysis, the most important priors used in this thesis are represented below.

1.5.1 Priors on the DNA substitution model

When analyzing DNA sequences in a Bayesian framework, one of the priors to set is the substitution model. In general, substitution models are assumptions about the nucleotide composition of the aligned sequences and the frequency of different mutation types. The simplest way to describe the distance between aligned molecular sequences is to calculate the proportion of deviating positions out of all sequence positions, considering only the number of differences, not the quality of them (p distance). However, p distance tends to oversimplify the sequence evolution, since it states that the probability of every substitution is the same. Therefore, more complicated evolutionary models have been developed for nucleotide sequences to describe the rate of fixed mutations. In these models, it is possible to take into consideration the possible differences in nucleotide frequencies, as well as set distinct probabilities for transitions (changes within purines and pyrimidines: A↔G and C↔T) and transversions (changes between purines and pyrimidines: A/G↔C/T).

The simplest substitution model by Jukes and Cantor (JC69) (Jukes and Cantor 1969) reflects the state where all bases appear at equal frequency and transitions and transversions emerge at the same rate (**Figure 8.** 'Some commonly used substitution models for DNA'). However, this is rarely the situation in real, naturally occurring populations and several extensions to the JC69 model have been created over decades. Among these are the Hasegawa, Kishino & Yano (HKY) model, which allows distinctive base frequencies and separate estimates for transitions and transversions (Hasegawa *et al.* 1985),

and the Tamura Nei (TN93) model, which takes the HKY model further and allows transitions to occur at different frequencies (Tamura and Nei 1993). The most complex model, the general time reversible (GTR) model, additionally distinguishes different rates for all possible transversion events (Tavaré 1986). In addition to the variability in the base and substitution frequencies, the proportion of invariant sites (pInv) (Fitch 1986; Shoemaker and Fitch 1989), and/or the among-site rate heterogeneity (+Γ) (Yang 1994) could be included in any of the models mentioned above. Rate variation across the sites is modelled with the gamma distribution and its shape parameter α ($\in [0, 1]$). Small values for α indicate strong heterogeneity, a few positions having a high substitution rate and others being basically invariant. Larger values suggest only weak variation within the substitution rates among the sequence positions. Bayesian inference allows setting prior distributions for each of the parameters included in the substitution model, such as nucleotide frequencies (π_A, π_C, π_G, π_T), transition to transversion rate (κ), shape parameter in gamma distribution (α), and proportion of invariant sites (pInv).

a)

	A	C	G	T
A		a	a	a
C	a		a	a
G	a	a		a
T	a	a	a	

b)

	A	C	G	T
A		$a\pi_C$	$b\pi_G$	$a\pi_T$
C	$a\pi_A$		$a\pi_G$	$b\pi_T$
G	$b\pi_A$	$a\pi_C$		$a\pi_T$
T	$a\pi_A$	$b\pi_C$	$a\pi_G$	

c)

	A	C	G	T
A		$c\pi_C$	$a\pi_G$	$c\pi_T$
C	$c\pi_A$		$c\pi_G$	$b\pi_T$
G	$a\pi_A$	$c\pi_C$		$c\pi_T$
T	$c\pi_A$	$b\pi_C$	$c\pi_G$	

d)

	A	C	G	T
A		$a\pi_C$	$b\pi_G$	$c\pi_T$
C	$a\pi_A$		$d\pi_G$	$e\pi_T$
G	$b\pi_A$	$d\pi_C$		$f\pi_T$
T	$c\pi_A$	$e\pi_C$	$f\pi_G$	

Figure 8 Some commonly used substitution models for DNA. a) JC69, b) HKY c) TN93 and d) GTR. π_A, π_C, π_G and π_T represents the nucleotide frequencies whereas a, b, c, d, e and f stand for transition/transversion probabilities.

In general, the more complicated the evolutionary process has been (i.e. the longer time has passed since the divergence), the more complex the model needed to explain the observed differences. Misspecification of the substitution model may have a huge impact on the phylogenetic inference (Posada and Crandall 2001) and therefore several methods have been developed to estimate the best-fit substitution model(s), such as PartitionFinder (Lanfear *et al.* 2012) and bModelTest (Bouckaert and Drummond 2017).

1.5.2 Priors on the tree model

While we might have relatively good foreknowledge of the prior distributions associated with the substitution model's parameters (see section 1.5.1 'Priors on the DNA substitution model'), specification of prior probabilities on different tree topologies describing the diversification process might be challenging. In Bayesian phylogenetics, models used as tree priors are

essentially functions that assign prior probabilities to distinct tree topologies, alongside determining the prior distributions of node times distributed along the tree. In Bayesian inference, there are two main approaches to specify tree priors: coalescent based models and models based on the birth-death process. In coalescent based models, the history within a population or species is traced backwards in time (Kingman 1982) and thus it is possible to detect the changes over time in the number of lineages. In contrast to the coalescent-based models, the underlying idea of birth-death processes is to model forward in time speciation ('birth') and extinction ('death') of lineages included in the phylogeny (Kendall 1948). Since particular birth-death models allow detection of changes in birth and death rates in a phylogenetic tree, as well as estimation of the underlying transmission process, these models have been used extensively, especially in viral epidemics. Furthermore, whereas coalescent-based models assume a small random sample drawn from a relatively large background population, the dataset used in birth-death based approaches can represent comparatively dense sampling. All the priors on the tree model used in this thesis are based on coalescent theory.

1.5.2.1 Coalescent based models

When the data consists of individuals from the same population or species, the history of a population could be outlined by tracing back the random coalescent events of pairs of lineages until the most recent common ancestor is obtained (Kingman 1982). This enables the estimation of the relationship between the population's genealogy and its demographic history. Since the coalescent theory was originally based on the Wright-Fisher model (Fisher 1930; Wright 1931), it makes simplified assumptions about the population structure. These include, for instance, panmixia, constant population size, and discrete generations. Quite often, however, these assumptions are violated. This in turn means that the population size estimated with the coalescent based model (i.e. effective population size, N_e) is almost without exception smaller than the actual census size of the population (N) (Wright 1931). In general, N_e describes the size of an ideal population, where there is an equal amount of drift per generation as in the population of interest. If the reproductive variation within the population increases, the effective population size decreases as the parental contribution to the next generation becomes more unequal. For humans, the effective population size has shown to be approximately one third of the census population size (Browning and Browning 2015).

In the case of mitochondria and the Y-chromosome, the effective population size is only a quarter of that of autosomes due to their haploid nature and gender-specific inheritance. In addition, differences in gender-specific generation times potentially have an impact on effective population size, as the gender with shorter generation time might experience more drift. For humans, it has been estimated that the generation time for males is 31

years and for women 25 years (Fenner 2005) which implies that N_e for mtDNA should be lower than that for the Y-chromosome, assuming 1:1 sex-ratio and equal reproducing success for both genders.

Effective population size along the genealogy can be outlined using traditional coalescent-based methods. It requires pre-determination of a so-called demographic model, which describes the change in the population size through time. These parametric demographic models include, for instance, constant population size, exponential growth, and logistic growth (for more details see Pybus *et al.* 2000; Drummond *et al.* 2005). As the demographic past of the population might be unknown or challenging to determine, nonparametric skyline plot models have been developed (Pybus *et al.* 2000; Drummond *et al.* 2005, see below). Since the skyline plot models estimate the past dynamics directly from the sequence data, no *a priori* knowledge of the population's demographic history is needed (Drummond *et al.* 2005).

1.5.2.2 Bayesian skyline plot model

In principle, skyline plot models proceed in two distinct steps: first determining the tree topology with (relative) branch lengths and then estimating the population history based on the genealogy (see Ho and Shapiro 2011). After resolving the genealogy, the mean population size (N) for each coalescent interval can be estimated with

$$N_i = \frac{\gamma_i i(i-1)}{2}, \quad (2)$$

where γ_i denotes the size of the coalescent interval and i denotes the number of lineages in a given interval. In contrast to many other skyline plot methods, the Bayesian skyline plot (BSP) estimates both the genealogy and the past population dynamics simultaneously based on sequence data (Drummond *et al.* 2005). This co-estimation reduces the effect of possible bias introduced by the inaccuracy in the tree topology reconstruction in the following step. Furthermore, the BSP takes into account the uncertainties in the tree topology and branch lengths, as well as the ambiguity in the reconstruction of the demographic history based on the genealogy (Drummond *et al.* 2005). This combined uncertainty is represented by highest posterior density (HPD) intervals (Drummond *et al.* 2005).

1.5.3 Priors on the evolutionary clock model

The hypothesis of the molecular clock was first introduced in the 1960s when it was observed that certain protein sequences appeared to evolve in a clock-like manner among particular mammalian species (Zuckerkandl and Paulig

1962; Sarich and Wilson 1966; Sarich and Wilson 1967) (See also section 1.5.4 'Calibrating the molecular clock'). Consequently, the molecular clock hypothesis suggests that the genetic difference between any two taxa is proportional to the time since they last shared a common ancestor. Soon after the introduction of the molecular clock hypothesis, Motoo Kimura proposed that most of the polymorphisms are selectively neutral, since mutations increasing fitness are rare and deleterious mutations are removed from the population by natural selection (Kimura 1968). The neutral theory states that most of the variation observed within and between populations is due to random genetic drift of neutral alleles. Moreover, the rate at which these neutral mutations are being fixed is approximately equal to the per-individual mutation rate.

Although Zuckerkandl's and Pauling's hypothesis of the molecular clock together with Kimura's neutral theory of evolution revolutionized the field of phylogenetics by allowing the inclusion of temporal dimension into the analyses, these theories have also been criticized. Simplified assumptions of this so-called 'strict molecular clock' have shown to be unrealistic as many taxa empirical studies have demonstrated inconsistency in molecular rates both over time and among lineages (for review see for example Bromham and Penny 2003; García-Moreno 2004; Kumar 2005; Bromham 2009). For the between-species deviation, several factors have been proposed as a source of variability, such as differences in generation times and metabolic rates (see García-Moreno 2004 and references therein). Furthermore, differences have been considered to arise due to evolutionary processes, such as selection and drift (see Bromham and Penny 2003 and references therein). In addition, nearly neutral theory (Ohta 1987; Ohta 2002) suggests that the effective population size also plays an important role, since alleles with neutral and nearly neutral selective pressure are expected to fixate faster in a small population due to genetic drift (for review see Bromham and Penny 2003).

Since the utilization of the unsuitable molecular clock model might bias not only the divergence date estimations (see for example Yoder and Yang 2000) but also the topology of the phylogenetic tree (Ho and Jermiin 2004), relaxed clock models allowing rate variation between the lineages have been developed (for example Thorne *et al.* 1998; Aris-Brosou and Yang 2002). The majority of relaxed molecular clock models assume that within a phylogenetic tree, branch rates are autocorrelated, meaning that the rate for each branch is modelled as a function of the parent branches' rate (Thorne *et al.* 1998; Aris-Brosou and Yang 2002).

The major drawback of relaxed clock models, however, is that the user has to determine the topology of the tree prior to the analysis, which might be challenging given that many equally probable phylogenies might exist (see Drummond *et al.* 2006). Therefore, 'relaxed phylogenetic' methods have been established where both the tree topology and divergence times are estimated in parallel (Rannala and Cranston 2005; Drummond *et al.* 2006). In addition, state-of-the-art relaxed clock models, implemented in the Bayesian

framework, do not assume autocorrelation between parent and child branch rates. Instead, the rate is independently drawn for each branch from a predetermined distribution, such as from lognormal or exponential distribution (Drummond *et al.* 2006).

1.5.4 Calibrating the molecular clock

Incorporating a molecular clock model into phylogenetic analyses will yield relative time estimates for divergence events in the phylogenetic tree, meaning that we are able to infer proportional time since the taxa shared a common ancestor. However, it is impossible to distinguish the absolute timescale of divergence events based purely on the accumulated sequence differences. Thus, regardless of the phylogenetic method used, all molecular clock models have to be calibrated with external time related information to be able to obtain concrete dates. Once calibration is included, it is subsequently possible to estimate, for example, the mutation rate per time unit (year or generation) from the data. (For review see for example Bromham and Penny 2003 Box 1; Ho 2015).

Several different techniques can be used to investigate the absolute timing of divergence events. Some commonly used methods are calibration of an internal or a terminal node of the phylogenetic tree or utilization of fixed rate estimate. As a node calibration, one can use, for instance, information from fossil records, such as in Zuckerkandl and Pauling 1962 (see above), where temporal information was gained by using a human-horse divergence estimated based on fossil data as a reference point. However, since the fossil record provides only minimum dates for the taxa divergence, age estimates for the most recent common ancestor might be underestimated. In addition, a dated geological event which has subsequently led to population divergence can serve as a source for a node calibration. However, population-divergence events do not necessarily temporally correspond to genetic divergence (Edwards and Beerli 2000), which can lead to overestimation of molecular rates and thus other methods are recommended. For a general review of different calibration methods see Box 1 in Ho *et al.* 2011.

For the terminal nodes, sometimes also referred to as the tips of the tree, the independent temporal calibration can be obtained from the direct dating of the sample(s) included in the analysis. For rapidly evolving short-lived species, such as bacteria, this dating could be the exact date of the sampling. For species with notably lower mutation rates, such as humans, radiocarbon (^{14}C) dated ancient samples (estimating time of death) can be used as a tip calibration to infer the absolute timescale. Even though tip calibration does not make any assumptions about the simultaneity of genetic and population divergences, the resulting molecular rate estimate might be biased, especially if there is uneven phylo-temporal clustering of the samples analyzed (for review see Rieux and Balloux 2016; Tong *et al.* 2018). However, incorporation

of time-structured sequence data with known radiocarbon ages is currently a commonly used approach in the field of ancient DNA studies.

For biological sequence data, the genetic differences between taxa can also be converted into absolute units of time by using a 'fixed' molecular rate. This means that the evolutionary rate calculated for one set of taxa is further extrapolated to another set of taxa (also known as 'universal clock', see for example García-Moreno 2004). The rate used might originate, for instance, from another study population of the same species or alternatively from evolutionarily closely related species. However, as shown for example in Weird & Schluter 2008, even evolutionarily closely related species can harbor notable variation in molecular rates. Similarly, for some cases within-species rates have shown to be heavily dependent on the dataset used and the timespan in question (see section 1.5.5 'Variation in the molecular rates - Mitochondrial point of view'). Due to this substantial variation between the different molecular rates, utilization of a fixed rate as the only source for calibration should be carefully considered.

A phylogenetic tree can contain several distinct calibration points, for instance both internal and terminal node calibrations. Furthermore, uncertainties in the node ages could be considered by specifying a prior distribution for the node age rather than by assigning a point estimate on it (Ho and Phillips 2009). For tip-calibrated trees, this means that the analysis can include several heterochronically sampled sequences and for each of the tip-ages, the probability distributions are also taken into account in the model. For instance, for ancient human samples, the probability densities of the radiocarbon dates can be incorporated into the analysis.

For human mtDNA, tip calibration has shown to yield more consistent estimates of molecular rates than the internal node calibration, which usually harbors larger uncertainty in the prior distributions (Rieux *et al.* 2014). Since inappropriate calibration might bias not only the evolutionary timescale but also the Bayesian estimates of the parameters that are dependent on the molecular rate, such as effective population size, careful consideration of the calibrations utilized is needed.

1.5.5 Variation in the molecular rates – Mitochondrial point of view

When considering the molecular rates, it is important to distinguish between the four types of rates discussed throughout this thesis (definitions according to Ho and Larson 2006): the mutation rate, pedigree rate, substitution rate and phylogenetic rate. The mutation rate characterizes the probability of spontaneous mutation event(s) in the genomic sequence, whereas pedigree rate is an estimate of mutation frequency determined based on the successive generations. The substitution rate represents the rate at which mutations are being fixed in a population and could be inferred, for example, from time-calibrated phylogenetic analyses. The last rate, the phylogenetic rate, is estimated based on the genetic differences observed between distinct species.

For humans, several studies have shown that pedigree-based mitochondrial molecular rates are roughly three times higher than the estimates obtained from within-human phylogenetic analyses (i.e. substitution rates) (Stoneking *et al.* 1992; Forster *et al.* 1996; Henn *et al.* 2009). Further, both of these rate estimates exceed the phylogenetic rates for mtDNA determined based on human-chimp divergence (Forster *et al.* 1996; Parsons *et al.* 1997; Howell *et al.* 2003; Santos *et al.* 2005). Interestingly, approaches using aDNA as a tip calibration seem to produce rates falling between the short-term pedigree-based estimates and long-term substitution rates (Ho, Shapiro, *et al.* 2007; Ho, Kolokotronis, *et al.* 2007).

A similar pattern of rate inconsistency emerges for the mtDNA across a variety of taxonomic groups and this has led to a theory of time dependency of molecular rates (Ho *et al.* 2005; Ho and Larson 2006; Ho, Shapiro, *et al.* 2007; Ho, Kolokotronis, *et al.* 2007; Ho *et al.* 2011 and Ho *et al.* 2015). The theory proposes that the observed inconsistency in molecular rates arises since the rates measured on different time scales reflect different biological processes, with short-term rates representing all spontaneous mutations (excluding lethal ones) and long-term rates considered to mirror the rate of mutation fixation (for discussion see Ho *et al.* 2011; Ho *et al.* 2015). Although the exact factor behind the variation still remains uncertain, the theory suggests selection as being the strongest candidate, possibly accompanied by genetic drift (See for example Ho and Larson Box 1).

In principle, newly emerged alleles can be fixed either through positive selection or due to random genetic drift. Regarding the deleterious mutations, it is evident that the longer the temporal period, the longer time natural selection has to remove these polymorphisms from the population and thus advantageous mutation(s) might be fixed. However, in the case of slightly deleterious mutations it has been shown that these nearly neutral mutations might become fixed, particularly in small populations due to a random fluctuation in the allele frequencies between successive generations (Ohta and Kimura 1971; Ohta 1987; Ohta 2002). Furthermore, when considering strictly neutral mutations, the only evolutionary process that might lead to fixation of the mutation is genetic drift, since they are unaffected by selection. Thus, according to the theory of nearly neutral evolution, the changing interplay between selection and drift over time introduces variation into the rates of molecular evolution (For review see Bromham and Penny 2003 and especially Box 4 therein). In short, it could be hypothesized that variation in the molecular rates is not driven only by the rate of new mutations arising but also by the balance between selection and drift, which are in turn influenced by the length of the time period under scrutiny and by the effective population size and its fluctuations between the generations.

Despite evidence having been observed within numerous species which supports the time-dependency of molecular rates, especially for mitochondrial DNA (see Ho *et al.* 2015), the theory is not unanimously accepted. It has been stated that the majority of the observed deviation is due to artefacts in the data

analysis, such as calibration errors or model misspecification (Emerson 2007; Emerson and Hickerson 2015). Plausible biases arising from these factors are indeed also acknowledged by Ho et al. (see Ho *et al.* 2011 and 2015) and can be controlled, at least to some extent, for example by selecting the calibration information carefully and by thorough model testing. For a comprehensive discussion concerning the opposing views between Ho *et al.* and Emerson *et al.*, please see both references: Emerson and Hickerson 2015 and Ho *et al.* 2015.

Rate variation is also detected within the mitochondrial genome: the non-coding control region evolves faster than the coding region, where natural selection is stronger. Several different estimates for substitution rates have been proposed along the mitochondrial sequence, but most widely used are 1.79×10^{-7} transitions/bp/year for HVR1 (nucleotides 16,090–16,365) (Forster *et al.* 1996), 2.28×10^{-7} mutations/bp/year for HVR2 (nucleotides 68–263) (Soares *et al.* 2009) and 1.69×10^{-8} mutations/bp/year for coding region (Friedlaender *et al.* 2005). Further, the control region in particular contains some additional positions which tend to be mutated more frequently than rest of the mitochondrial genome (see Soares *et al.* 2009 and references therein). High mutational frequency in these so-called mutation hotspots might result in homoplasy, where haplotypes having the same mutation(s) appear to be evolutionarily related, though lineages have gained mutation(s) independently. To avoid false classification of sequences, the mutation hotspots are typically omitted from the population-level comparisons.

While several studies have focused on the differences in the mtDNA molecular rates obtained by the distinctive calibration methods and further studies have focused on the molecular rate differences along the human mitochondrial genome, no systematic comparison of molecular rates between the mitochondrial lineages has been performed.

2 POPULATION HISTORY OF EUROPE – GENETIC OVERVIEW

During the past few decades, the population history of Europe has been inferred by exploiting the ancient DNA (aDNA) extracted from archaeological human remains, in addition to archaeological evidence and modern genetic data. The rapidly growing field of aDNA research has shed more light onto several archaeological hypotheses, although various questions remain unanswered and more multidisciplinary research is needed. Nevertheless, using aDNA it could be proposed that the genetic past of Europe encompasses several population turnovers and migrations, such as the spread of farming-based populations from the Near East (starting ~10,000 years ago) and the extensive westward migration from the Pontic-Caspian steppe into Europe approximately 5,000 years before present (ybp).

Anatomically modern humans have occupied continental Europe permanently since the initial dispersal of foraging hunter-gatherer populations into Europe occurring around ~45,000 years ago. However, the genetic contribution of these archaic populations to the contemporary gene pool have shown to be modest (Fu *et al.* 2016) and the earliest individuals with identifiable genetic continuation to the present-day Europeans existed notably later, approximately 36,000–39,000 ybp (Seguin-Orlando *et al.* 2014; Fu *et al.* 2016). Climatologically this period was characterized by global cooling, reaching its maximum around 27,000–19,000 ybp, during which Northern Eurasia became largely covered by an ice sheet. The following deglaciation period resulted in the gradual re-dispersal of hunter-gatherers from the refugial regions and Europe became primarily inhabited by the 'Western hunter-gatherers' (WHG) (Lazaridis *et al.* 2014; Fu *et al.* 2016; Mathieson *et al.* 2018; Mittnik *et al.* 2018) (see also Lazaridis 2018 and references therein). Whereas WHG dominated in Southern, Western, and Central Europe, parts of European Russia were populated by the 'Eastern hunter-gatherers' (EHG), carrying both WHG and archaic Siberian heritage (Haak *et al.* 2015; Mathieson *et al.* 2015). Subsequently, both these ancestral origins, WHG and EHG, contributed to the genomic composition of Scandinavian and Baltic hunter-gatherer populations (Lazaridis *et al.* 2014; Skoglund *et al.* 2014; Jones *et al.* 2017; Saag *et al.* 2017; Günther *et al.* 2018; Mittnik *et al.* 2018).

Animal domestication and cereal cultivation, accompanied by other innovations such as ceramics, originated in the Near East around 11,500 years ago. Subsequently, this 'Neolithic package' started to disperse gradually into Europe, although its cultural and genetic impact varied across the continent. In Southern and Central Europe along with Scandinavia, the Neolithic transmission included an influx of new genetic material, most likely of Anatolian origin (Mathieson *et al.* 2015; Hofmanová *et al.* 2016; Lazaridis *et al.* 2016; Omrak *et al.* 2016), supporting a demic diffusion model (see Fort

2015). In contrast, for instance in the Baltic, the shift from hunting and gathering to a farming-based way of life happened thousands of years later, and most likely did not involve dissemination of the Anatolian genetic component (Jones *et al.* 2017; Saag *et al.* 2017).

During the late Neolithic and early Bronze Age, a large-scale population migration across Europe occurred originating from the East European Steppe, starting approximately 5,000 years before present (ybp) (Allentoft *et al.* 2015; Haak *et al.* 2015). This dispersal, mainly driven by individuals associated with Yamnaya culture, brought along a new genetic component which largely replaced the previously prevalent Neolithic ancestry (Allentoft *et al.* 2015; Haak *et al.* 2015). Population expansion carrying this 'Steppe ancestry' therefore contributed vastly to the genomic composition of many parallel and subsequent populations, and has even been related to the spread of Indo-European languages (Haak *et al.* 2015).

Simultaneously, during the late Neolithic (~4,800–2,200 ybp) a new material entity called Corded Ware culture (CWC) emerged. Geographically, the spread of CWC comprised a vast area covering the majority of Central and Eastern Europe as well as the southern parts of Fennoscandia. Based on archaeological evidence, CWC has been influenced by both the preceding Neolithic cultures and Yamnaya (for review see Kristiansen *et al.* 2017). The latter is further supported by the notable amount of Yamnaya related ancestry among the Corded Ware people (Haak *et al.* 2015). Traditionally, CWC groups have been considered to be mobile herders, but evidence of local cereal cultivation supports a mixed subsistence economy (see Sjögren *et al.* 2016 and references therein). On the northern fringes of Europe, such as in the Baltics, the onset of intensive cultivation and animal domestication has been associated with the appearance of CWC (~4,800–4,000 ybp), genetically characterized by the arrival of Steppe ancestry (Jones *et al.* 2017; Saag *et al.* 2017). Additionally, Fennoscandia and the Eastern Baltics were later influenced by considerable Eastern gene flow arriving from Siberia no later than 3,500 years ago (Lamnidis *et al.* 2018; Saag *et al.* 2019; Sikora *et al.* 2019).

Even though Europe has experienced several migration waves and even population turnovers during its prehistory, most of these different genetic origins have left their signature on the subsequent genome pool. In general, Europeans are a complex admixture of various ancestral layers, with contemporary populations displaying these components in distinctive proportions.

3 POPULATION HISTORY OF EUROPE – MITOCHONDRIAL POINT OF VIEW

Ancient DNA studies have revealed that before the Last Glacial Maximum (LGM), the earliest European hunter-gatherers carried mitochondrial lineages belonging to both macrohaplogroups M and N, with lineages U2, U5 and U8 being particularly common (Krause *et al.* 2010; Fu *et al.* 2013; Posth *et al.* 2016). During the following post-LGM period, virtually only sublineages of N are present, suggesting a population bottleneck presumably driven by the environmental changes (Posth *et al.* 2016). At the same time as the Late Glacial population expansion, starting approximately 14,500 ybp, U5a and U5b became dominant lineages in Europe (Posth *et al.* 2016), albeit these haplogroups most likely evolved already during the LGM in refugial regions (Malyarchuk *et al.* 2010). The high prevalence of U5 sublineages also persisted during the Mesolithic Stone Age with U5a being mostly distributed in Northern and Eastern Europe and U5b typically among the Central and Southern European hunter-gatherers (Bramanti *et al.* 2009; Malmström *et al.* 2009; Hervella *et al.* 2012; Sánchez-Quinto *et al.* 2012; Skoglund *et al.* 2012; Bollongino *et al.* 2013; Fu *et al.* 2013; Der Sarkissian *et al.* 2013; Lazaridis *et al.* 2014; Skoglund *et al.* 2014; Haak *et al.* 2015; Malmström *et al.* 2015; Posth *et al.* 2016; Jones *et al.* 2017; Mathieson *et al.* 2018; Mittnik *et al.* 2018). In addition to U5, haplogroup U4 was widely spread during the Mesolithic, reaching its highest frequencies particularly in Scandinavia, the Baltics and Northwestern Russia (Malmström *et al.* 2009; Skoglund *et al.* 2012; Der Sarkissian *et al.* 2013; Skoglund *et al.* 2014; Malmström *et al.* 2015; Saag *et al.* 2017; Günther *et al.* 2018; Mittnik *et al.* 2018).

At the onset of the Neolithic period (~10,000 ybp), new mitochondrial haplogroups were introduced to Europe with the expansion of the farming populations from the Near East (Mathieson *et al.* 2015; Hofmanová *et al.* 2016; Lazaridis *et al.* 2016; Omrak *et al.* 2016). These lineages, including mostly haplogroups H, HV, J, K, N1a and T2, partially replaced the preceding U lineages (Haak *et al.* 2005; Bramanti *et al.* 2009; Brotherton *et al.* 2013), although the presence of U in contemporary Europeans reveals that hunter-gatherers were eventually assimilated into the arriving farmer populations. However, only modest local maternal genetic continuity has been observed between the early and middle/late Neolithic farmers in Europe (Brotherton *et al.* 2013), which suggests that several migration waves altering the gene pool took place during the Neolithic period.

Even if the prevalence of U decreased considerably during the Neolithic, lineages U5a and U4 subsequently experienced re-expansion during the late Neolithic and Bronze Age, from 5,000 ybp onwards (Keyser *et al.* 2009; Wilde *et al.* 2014; Allentoft *et al.* 2015; Haak *et al.* 2015; Mathieson *et al.* 2015; Pilipenko *et al.* 2015; Mittnik *et al.* 2018). For instance, U5a displays up to

20% frequency in populations associated with Yamnaya culture, expanding westwards from the Pontic Steppe into Europe (Keyser *et al.* 2009; Wilde *et al.* 2014; Allentoft *et al.* 2015; Haak *et al.* 2015; Mathieson *et al.* 2015; Pilipenko *et al.* 2015; Mittnik *et al.* 2018). Signs of this reappearance are also visible during the later eras; U5a is typical in Northeastern Europe during the Early Metal period (~3,500 ybp) (Der Sarkissian *et al.* 2013) and frequent among the Iron Age Scythians from Southeastern Europe (~2,700–2,200 ybp) (Juras *et al.* 2017).

Owing to the abundance of U during the Mesolithic and the later emergence and spread of lineages H, J, K and T during the Neolithic, haplogroup U is traditionally considered to be a 'hunter-gatherer haplogroup', whereas the latter haplogroups are jointly classified as 'farmer haplogroups'. Naturally, an individual's mode of subsistence cannot be inferred based on the mitochondrial haplogroup; this is just a convention used to refer to populations where each haplogroup has been most typical. If one is interested in the potential food sources of ancient individuals, stable isotope analyses are required.

Various aDNA studies have shed light on the past of mitochondrial lineages in Europe. Since the dispersal of the anatomically modern human from Africa to other continents, the European mitochondrial gene pool has been influenced by various processes, in particular by numerous migration events and genetic drift (for review see Torroni *et al.* 2006). Independent migration events have brought new mtDNA lineages into Europe and haplogroup frequencies have fluctuated through millennia. Furthermore, it has been proposed that the climate has also shaped the geographical variation of mitochondrial haplogroups through adaptive selection (Mishmar *et al.* 2003), although this hasn't been supported by later studies (Moilanen *et al.* 2003; Elson *et al.* 2004). However, since the majority of haplogroups observed in the prehistoric era are still present among contemporary populations, these different ancestries have contributed to the present day European mitochondrial gene pool.

4 POPULATION HISTORY OF FINLAND

4.1 Archaeological background of prehistoric Finland

The earliest postglacial signals of human activity in the geographical area of present-day Finland date back approximately 10,900 years (see Haggren *et al.* 2015 pp. 26 and references therein). During that time, Finland was partially covered by the ice sheet and the southwestern areas resided largely below sea level. Presumably, the first settlers arrived from south, southeast and additionally along the ice-free northern coast. Similar to all the other human groups around Europe at the time, food was obtained by hunting and fishing, as well as by consuming wild plants.

The emergence of the first material cultures occurred during the Early Combed Ware period starting 7,300 ybp, with Sperrings ceramics prevalent in Southern Finland and Säräisniemi 1 ceramics mainly in the northern areas (Pesonen *et al.* 2012). In Finland, the appearance of ceramics defines the transition from Mesolithic to Neolithic Stone Age (10,900–7,200 ybp and 7,200–1,800 ybp, respectively). This is in contrast to many other parts of Europe, where the onset of the Neolithic period is conventionally associated with the arrival of the sedentary farming-based lifestyle.

The Early Combed Ware period was followed by the era of Typical Combed Ware (TCW), which spread to almost the entire geographical area of present-day Finland approximately ~5,900–5,500 years ago (Carpelan 1999; Tallavaara *et al.* 2010). The dispersion of TCW, spearing from the East, was most likely accompanied by the arrival of new human groups and thus also genes (Haggren *et al.* 2015 pp. 60). Interestingly, this period is characterized by high annual temperature and ecosystem net productivity coinciding with the Stone Age population peak, suggested by the abundance of archaeological material (Oinonen *et al.* 2010; Tallavaara and Seppä 2012; Oinonen *et al.* 2014; Tallavaara and Pesonen 2020). The highest population level was attained during TCW approximately 5,800 ybp, but the population size started to decline shortly after, simultaneously with the late-Holocene cooling (Tallavaara *et al.* 2010; Tallavaara and Pesonen 2020). This decrease continued until the lowest population densities were reached somewhere around 4,100–3,800 years ago (Tallavaara and Seppä 2012; Sundell *et al.* 2014; Tallavaara and Pesonen 2020). Temporally parallel to the final stages of TCW culture, Early Asbestos Ware and related cultures emerged in Eastern Finland (~6,600–4,500 ybp) (Oinonen *et al.* 2014) whereas the southwestern parts of the country were influenced by Corded Ware culture somewhat later (~4,800–4,300 ybp) (Carpelan 1999; Oinonen *et al.* 2010; Haggren *et al.* 2015). These events have contributed to creating a cultural and genetic East-West distinction within Finland (for cultural distinction see for example

Sarmela 2009 'Finnish Folklore Atlas: Ethic Culture of Finland' and for genetic East-West distinction see section 4.2. 'Population genetics of contemporary Finns').

During the subsequent Early Metal period (3,800–1,700 ybp), geographical differences were further seen in material cultures. In principle, Finland was divided into coastal and inland regions with distinct cultural influences. Along the coast, during period 3,700–2,500 ybp - also known as Bronze Age - the archaeological material shows notable Scandinavian influence, whereas inland areas show mainly Eastern features (see Haggren *et al.* 2015 pp.129–130). However, migration and particularly trading occurred between these two cultural entities over the Baltic Sea to Scandinavia and even to some extent to the Volga Region (see Haggren *et al.* 2015 pp.129–130, 172–173, 190, 210–211).

In Finland, the Iron Age started around 2,500 years ago and lasted until ~1,300 AD (*Anno Domini*). The first iron artefacts most likely arrived from two directions: from somewhere around the Baltic Sea to the coastal area and from all the way from Eastern Russia to the inland, although local production of iron was established soon after (Haggren *et al.* 2015 pp. 217–219). Even though the spatial differences in Eastern and Scandinavian influences are still apparent, intensive fur trading took place within the country and even extended abroad, particularly during the Viking Age and Crusade period (800–1,200 AD).

Even though occasional evidence of small-scale farming and animal husbandry exists from the Neolithic period and the Bronze Age in Finland (Alenius *et al.* 2013; Bläuer and Kantanen 2013; Cramp *et al.* 2014; Vanhanen *et al.* 2019), cultivation only became more intensive in the beginning of the common era (Haggren *et al.* 2015 pp. 227–228, 273, 304–305; Lahtinen *et al.* 2017). The farming prior to the Iron Age had most likely been mainly slash-and-burn agriculture, which was practiced alongside hunting and fishing (Haggren *et al.* 2015, pp. 132–133, 210–211, 227–228). In the southwestern areas, slash-and-burn farming was gradually diminishing already during the Iron Age, whereas in Eastern and Northern Finland it remained frequently practiced until the 19th century (Haggren *et al.* 2015, pp. 411, 471). The end of the Iron Age - 1,200 AD in the West and 1,300 AD in the East - marks the transition from prehistory to Medieval (1,200–1,525 AD), followed by modern times (1,520 AD onwards).

In conclusion, the territory of today's Finland has experienced several cultural influences during its prehistory. Over thousands of years, new cultural entities have arrived, particularly from the west, south and east. Influences have been distinctive within the country: the southwestern coastal areas presumably had more connections to Scandinavia whereas in the inland areas the Eastern origin was more pronounced. Potentially, at least to some extent, new human groups also arrived along with the spread of cultures. In addition, the population size underwent several fluctuations, most likely driven by the environmental changes.

4.2 Genetic background of contemporary Finns

The genetic structure of present-day Finns has been the target of interest for several decades, starting from the 1970s, when a notable genetic substructure in blood groups and polymorphic serum proteins was observed within Finland (Nevanlinna 1972). The same pattern has later also been detected with other genetic markers. Additionally, it has been proven numerous times that contemporary Finns differ genetically from other European populations. This genetic uniqueness of Finns has been interpreted to be a consequence of the inhabitation history influenced by a series of founder effects, geographical isolation, genetic drift and population bottleneck(s), which has been supported by genome-wide studies (Varilo *et al.* 2000; Varilo *et al.* 2003; Jakkula *et al.* 2008; Chheda *et al.* 2016). A well-known indication of this peculiar genetic composition is Finnish disease heritage (FDH), which consists of 36 diseases commonly found in Finland but rare or absent in other countries (Norio *et al.* 1973; Norio 2003; see also www.findis.org). In contrast, many inherited disorders relatively common in other European countries, such as cystic fibrosis, are rare in Finland. The majority of monogenic diseases included in FDH follow an autosomal recessive mode of inheritance. Single founder mutations account for around 70–100% of the disease cases (Norio 2003; see also www.findis.org).

Early population genetic studies on autosomal data propose that contemporary Finns are genetic outliers compared to many European populations (Lao *et al.* 2008; Nelis *et al.* 2009) but later results have shown that Finns are instead part of the genetic continuity between mainland Europe and Uralic-speaking populations from Siberia (Tambets *et al.* 2018). Genome-wide data has also revealed that there is clear genetic division between East and West (Hannelius *et al.* 2008; Salmela *et al.* 2008; Nelis *et al.* 2009; Kerminen *et al.* 2017), with Western Finns resembling other Europeans, such as Swedes and Estonians, (Salmela *et al.* 2008; Nelis *et al.* 2009) and Eastern Finns exhibiting a higher amount of Asian genetical contribution to their genomes (Salmela *et al.* 2008; Tambets *et al.* 2018).

The Y-chromosomes of Finnish males mainly belong to the same haplogroups as observed in other Western Eurasians, such as N-M46 (N1c1)[1]; I-DF29 (I1a)[2]; R-M420 (R1a) and R-M343 (R1b) (Lahermo *et al.* 1999;

1 For the Y-chromosomal haplogroup N the following nomenclatures are used parallel:
(Defining mutation / ISOGG 2016 / ISOGG 2019 / Ilumäe 2016):
N-M231 / N / N / N; N-M46/ N1c1 / N1a1 / N3; N-L708 / N1c1a1 / N1a1a1a / N3a;
N-VL29 / N1c1a1a1a1 / N1a1a1a1a1a / N3a3; N-Z1936 / N1c1a1a1b / N1a1a1a1a2 / N3a4.

2 For the Y-chromosomal haplogroups I and R the following nomenclatures are used parallel:
(Defining mutation / ISOGG 2016 / ISOGG 2019):
I-M170 / I; I-M253 / I1; I-DF29 / I1a; R-M207 / R; R-M173 / R1; R-M420 / R1a; R-M459 / R1a1;
R-M343 / R1b.

Lappalainen *et al.* 2006; Palo *et al.* 2009), however the Y-chromosomal diversity is reduced and the frequencies of the haplogroups differ substantially when compared to many European populations (Sajantila *et al.* 1996; Lahermo *et al.* 1999; Lappalainen *et al.* 2006; Palo *et al.* 2009). Whereas in most European populations the most common Y-chromosomal haplogroups are I-DF29; R-M420 and R-M343 (see **Figure 7**), Finns show an exceptionally high proportion of lineage N-M46, reaching frequencies of 60 % (Lahermo *et al.* 1999; Lappalainen *et al.* 2006; Lappalainen *et al.* 2008). On a subhaplogroup level, the most common N-M46 sublineages in Finland are N-VL29 and N-Z1936 (Ilumäe *et al.* 2016; data from Altshuler *et al.* 2010), which are abundant also in many other geographically close Uralic-speaking populations, such as the Saami and the Estonians, and also prevalent in Eastern Europe (Ilumäe *et al.* 2016). All of these findings suggest a strong Eastern contribution to the Finnish Y-chromosomal gene pool (Kittles *et al.* 1999; Lahermo *et al.* 1999; Lappalainen *et al.* 2006).

Moreover, within-country division is also visible in the male lineages: in the East the Y-chromosomal diversity is more constricted and the Eastern affinity is more apparent than in Western and Southern Finland (Kittles *et al.* 1998; Lahermo *et al.* 1999; Lappalainen *et al.* 2006; Palo *et al.* 2007; Palo *et al.* 2009). The frequency of lineage N-M46 is as high as 71% in Eastern Finland, whereas it is only around 40% in the West (Lappalainen *et al.* 2006, see also **Table 9** in results). In contrast, the most typical haplogroup in the South-West is I-DF29 (~40%, Lappalainen *et al.* 2006), especially common in Sweden (Tambets *et al.* 2004), indicating a stronger Scandinavian influence in Southwestern Finland (Lappalainen *et al.* 2006; Palo *et al.* 2009). This discrepancy proposes divergent male population histories in different regions; distinctive genetic influences have affected Eastern and Western Finland with unequal Eastern and Western contributions during the past.

In contrast to the autosomal and Y-chromosomal composition, the mitochondrial gene pool of Finns seems to resemble other European populations, as typical European haplogroups are prevalent such as H, I, J, K, T, U, V and W (Torroni *et al.* 1996) (**Figure 5**). As in most of the European populations, the most common maternal haplogroup among Finns is H, for which the frequency in Finland varies between 34 and 41 %, depending on the sampling size and location (Torroni *et al.* 1996; Meinilä *et al.* 2001; Hedman *et al.* 2007). By contrast, the frequency of the "hunter-gatherer" lineage U in Finland (up to 28%, Torroni *et al.* 1996; Meinilä *et al.* 2001; Hedman *et al.* 2007) is one of the highest in Western Eurasia. Only Saami and some indigenous populations from the Volga-Ural area, such as Chuvash and Komi, display higher prevalence (48%, 44% and 37%, respectively) (Sajantila *et al.* 1995; Dupuy and Olaisen 1996; Delghandi *et al.* 1998; Bermisheva *et al.* 2002; Tambets *et al.* 2004). In a departure from many other European populations, haplogroup Z, frequent in Asia, occurs in Finland at low frequencies (~2.5 %, Finnilä *et al.* 2001). However, no clear differences in mitochondrial diversity between Finns and other European populations have been detected (Sajantila

et al. 1995; Torroni *et al.* 1996; Meinilä *et al.* 2001; Hedman *et al.* 2007), though it must be noted that the majority of these studies have analyzed only HVR1 or HVR1+HVR2, meaning that a considerable amount of information is lost.

Apart from the autosomal and Y-chromosomal diversity, Finns seem relatively homogenous based on mtDNA haplotype and haplogroup frequencies (Lahermo *et al.* 1996) with only modest within-country differences detected (Hedman *et al.* 2007; Palo *et al.* 2009). Despite mitochondrial diversity being somewhat reduced in the North and the East compared with other parts of Finland, regional differences are considerably lower than that of the Y-chromosome (Palo *et al.* 2009). It has been proposed that these differences observed between regional patterns of mitochondrial and Y-chromosomal frequencies might result either from differences in mutation rates (Sajantila *et al.* 1996), sex biased migration, i.e. women's higher mobility (Sundell *et al.* 2010; Sundell *et al.* 2013), or from more pronounced genetic drift in Finnish males (Lappalainen *et al.* 2006).

While the population history of Europe has been studied comprehensively during recent decades using ancient DNA (see section 2. 'Population history of Europe - Genetic overview'), the geographical area of present-day Finland has been an exception until recently (Lamnidis *et al.* 2018; Översti *et al.* 2019; Sikora *et al.* 2019). This can be largely attributed to a lack of archaeological bone material: due to acidic soil, unburnt human skeletal remains older than ~2,000 years are extremely rare in Finland (Ahola *et al.* 2016), making it challenging to analyze the genetic diversity prior to the Iron Age. However, recent studies have shown that during the Iron Age (~1,500 ybp) the Siberian genetic influence was significantly more pronounced among the individuals from Western Central Finland than is seen in modern-day Finns (Lamnidis *et al.* 2018; Sikora *et al.* 2019). These individuals from the Levänluhta burial site resembled contemporary Saami populations, hinting that the Proto-Saami language was spoken in a geographically much wider area than the present-day linguistic area of Saami languages, as also suggested by the analysis of place names and loanwords in Finland (Aikio 2012; Häkkinen 2010). However, it should be noted, that these inferences are based on a small number of samples from one location. Additional aDNA research is needed, covering broader sampling both in time and space, to infer a more detailed picture of how the genetic variety visible among present-day Finns was formed.

AIMS OF THE STUDY

The aims of this thesis are to elucidate the following research questions via high-resolution analyses of complete mtDNA genomes in a Bayesian framework:

i) Is there a mitochondrial East-West divergence within Finland? (Study I)

ii) Do Finns differ from other European populations in terms of mtDNA? (Study II)

iii) What can mtDNA reveal about the arrival of agricultural populations to Finland? (Studies I & II)

iv) Is there variation in the molecular rates among the mitochondrial haplogroups? (Study III)

5 MATERIALS

5.1 Mitochondrial DNA data

All mitochondrial datasets analysed in this thesis were obtained either from previously published articles or through database searches. The databases utilized for data retrieval were: GenBank (https://www.ncbi.nlm.nih.gov/genbank/), the Human Mitochondrial database (HmtDB, https://www.hmtdb.uniba.it/, Clima *et al.* 2017), PhyloTree (van Oven and Kayser 2009) and the Ancient mtDNA database (AmtDB, https://amtdb.org/, Ehler *et al.* 2019). Additionally, data was collected from the 1000 Genomes project (1000_Genomes_Project_Consortium *et al.* 2012). Analyses included either HVR1+HVR2 and complete sequences (study I) or only complete mitochondrial genomes (studies II and III) (**Table 1**).

Table 1. Mitochondrial DNA data used in studies I–III. HVR1 and HVR2 included sequence positions 16,024–16,385 and 73–340, respectively. Unless stated otherwise, sequences are obtained from contemporary populations.

Study	Sample size	Mitochondrial sequence	Reference
Study I	832	HVR1+2	Palo *et al.* 2009
	367	Complete	Multiple sources, see supplementary table S1 in study I
Study II	843	Complete	Multiple sources, see table 1 in study II
Study III	234	Complete (ancient)	Multiple sources, see supplementary table S1 in study III
	301	Complete	PhyloTree v17 (van Oven and Kayser 2009), see supplementary table S2 in study III

In study I, two types of Finnish mtDNA data were used: 1) HVR1+2 sequences (Palo *et al.* 2009) and 2) complete sequences. Mitochondrial HVR1+2 data also included information about the donors' place of residence, enabling regional comparisons within Finland. In study I, individuals were divided into thirteen subpopulations corresponding to the provinces in Finland (see **Figure 9.** 'Map of Finland and 13 subpopulations').

Figure 9 Mitochondrial HVR1+2 and Y-STR data were divided into 13 subpopulations according to Palo *et al.* 2009. Geographically the division corresponds to former Finnish administrational provinces, except LMO, which is part of the Vaasa province, but populated by a Swedish-speaking community. Southern and western provinces: Åland (AL), Turku (TU), Uusimaa (UU), Häme (HA), Vaasa (VA) and Larsmo (LMO). Eastern and Northern provinces: Mikkeli (MI), Kymi (KY), Central Finland (CF), Kuopio (KU), Northern Carelia (NC), Oulu (OU) and Lapland (LA). (Figure from *PLoS ONE*, CC BY 4.0 license).

For study II, complete contemporary mitochondrial genomes from Finland (N=843) were obtained from various sources, such as from databases and from published articles. For more details, see **Table 1** and supplementary table S1 in study II.

Since the aim of study III was to estimate substitution rate variation between and within mitochondrial haplogroups, both ancient and modern sequences were used. Sublineages of haplogroup U (U2, U4, U5a and U5b) were chosen as a study case because it is one of the oldest haplogroups among Europeans – the earliest individuals belonging to U date back approximately 38,000 years (Krause *et al.* 2010). U was also chosen since there were plentiful complete mitochondrial genomes available from radiocarbon dated ancient individuals. Further, it is well known that different sublineages of U have experienced different demographic pasts (see section 3. 'Population history of Europe - Mitochondrial point of view'). Complete ancient mitochondrial genomes and associated metadata were obtained from the Ancient mtDNA database (Ehler *et al.* 2019, https://amtdb.org/) and from previously published articles (see supplementary table S1 in study III). With regards to the ancient data, only those samples which had been radiocarbon (^{14}C) dated were chosen, as the purpose was to incorporate the absolute time scale into the analysis by tip calibration. For consistency, the ^{14}C-dates of ancient samples

were calibrated with Oxcal 4.3 (Bronk Ramsey 2009) using the IntCal curve 13 (Reimer *et al.* 2013). To avoid possible bias introduced by missing data, samples containing more than 10% of missing nucleotides were discarded.

In addition to the substitution rate variation *between* the subhaplogroups of U, we were interested to see how the utilization of different datasets *within* the subhaplogroup affects the molecular rates. Therefore, contemporary mtDNA sequences were collected from PhyloTree v17 (van Oven and Kayser 2009). From the PhyloTree database, one mitochondrial haplotype was chosen from each subhaplogroup, such as U5b1b2 and U5b1b2a. Additionally, two Paleolithic samples belonging to haplogroup R were chosen to represent an outgroup in the phylogenetic analyses ('R-root'). These samples included Fumane 2 dating back 39,805 calibrated ybp (calYBP) (GenBank ID: KP718913, Benazzi *et al.* 2015) and Ust'-Ishim dating back 45,050 calYBP (Fu *et al.* 2014).

In short, for each subhaplogroup of U (U2, U4, U5a and U5b), complete mitochondrial genomes were collected from ancient and contemporary individuals and in addition two ancient samples from haplogroup R were chosen to represent an outgroup (R-root). Based on these sets of samples, for each subhaplogroup of U (U2, U4, U5a and U5b), three separate analyses were carried out (**Figure 10**):

A) Only ancient sequences
B) Ancient and contemporary sequences
C) Ancient sequences, contemporary sequences and R-root

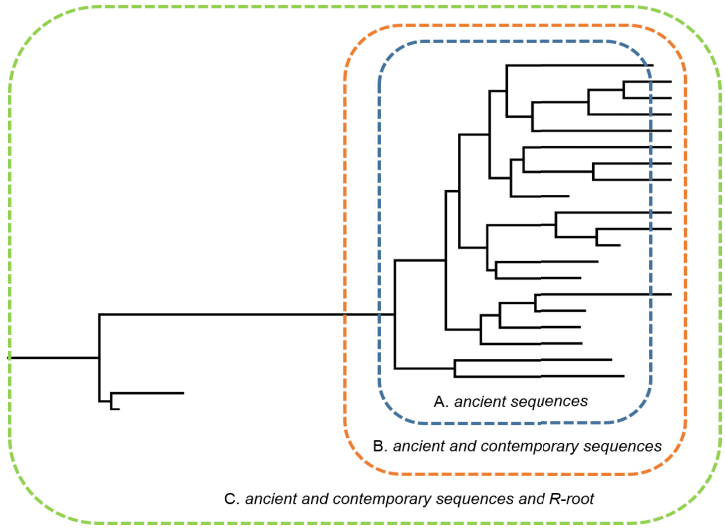

Figure 10 Schematic illustration of different datasets used in three distinct analyses for U2, U4, U5a and U5b. A) Blue = analysis containing only ancient sequences, B) Orange = analysis containing ancient and contemporary sequences and C) Green = analysis containing ancient sequences, contemporary sequences and R-root.

5.2 Y-chromosomal data

Study I included Y-chromosomal STR data from Finland from two different sources: in-house data (N=330 samples) and data obtained from the Family Tree DNA website (N=254 samples, https://www.familytreedna.com/). For the in-house data, the sample collection, DNA extraction, and Y-STR haplotype typing was performed as described in (Palo *et al.* 2009). The number of samples and the information regarding the Y-STR system and loci used are presented in **Table 2.**

Table 2. The Y-chromosomal data, kit (if available) and loci used in study I.

Study	**Sample size**	**Y-STR system used (loci included)**
Study I	330	AmpF1STR Yfiler kit (DYS456, DYS389I, DYS390, DYS389II, DYS458, DYS19, DYS385a/b, DYS393, DYS391, DYS439, DYS635, DYS392, Y-GATA-H4, DYS437, DYS438 and DYS448)
	254	N/a (DYS456, DYS389I, DYS390, DYS389II, DYS458, DYS19, DYS385a/b, DYS393, DYS391, DYS439, DYS392, Y-GATA-H4, DYS437, DYS438 and DYS448)

6 METHODS

6.1 Mitochondrial DNA

Several different methods were used to assess the original research questions (see aims of the study). **Table 3** represents an overview of the principal methods used to resolve the research questions i)–iv).

Table 3. Overview of the principal methods used to assess the research questions i)–iv). Unless stated otherwise, sequences analyzed were from contemporary populations. BSP=Bayesian skyline plot, hgs=haplogroups.

Research question	Analysis	Purpose of the analysis	mtDNA data used	Study
i)	Calculation of basic diversity indices	Determination of within Finland variation in mtDNA diversity	Finnish HVR1+2	Study I
i)	Linear regression	Estimation of the geographical variation within Finland for hunter-gatherer and farmer clusters	Finnish HVR1+2	Study I
ii)	BLAST searches	Comparison of Finnish dataset with GenBank database to evaluate if certain haplogroups are present only among Finns	Finnish complete sequences	Study II
ii)	Reconstructing Neighbor-Joining trees with MEGA	Estimation of the variation among Finn-characteristic haplogroups	Finnish complete sequences	Study II
ii)	Reconstructing phylogenetic trees with BEAST	Determination of divergence estimates for Finn-characteristic haplogroups	Finnish complete sequences	Study II
ii)	Reconstructing BSPs with BEAST	Determination of demographic past and comparison between Finn-characteristic and immigrant hgs	Finnish complete sequences	Study II
iii)	Reconstructing BSPs with BEAST	Determination of past population sizes for Finnish hunter-gatherer (U) and farmer (H) sequences and comparison of demographic past with European U and H	Finnish U and H coding region sequences	Study I
iv)	Reconstructing phylogenetic trees with BEAST	Estimation of molecular rate variation between mitochondrial lineages by utilizing radiocarbon dated ancient samples and contemporary sequences	Ancient and modern complete sequences from hgs U2, U4, U5a and U5b (mainly European origin)	Study III

6.1.1 Haplogroup determination

In study I, the mitochondrial haplogroups for the HVR1+2 and for the complete sequences were determined with HaploGrep (Kloss-Brandstätter *et al.* 2011), based on PhyloTree version 15 (van Oven and Kayser 2009).

Additionally, for the HVR1+2 data haplogroups were further confirmed by reconstructing unrooted maximum likelihood trees and visually inspecting the position of the samples in the phylogenetic tree. The maximum likelihood trees were built with MEGA v. 5.05 (Tamura *et al.* 2011), using the Tamura-Nei substitution model with the among-site rate heterogeneity (+Γ) with shape parameter α=0.7.

In study II, the mitochondrial haplogroups were assessed with two different methods: Haplofind (Vianello *et al.* 2013) and HaploGrep2 (Weissensteiner *et al.* 2016), based on PhyloTree versions 16 and 17 respectively. If these two methods yielded different subhaplogroups for a certain sample, HaploGrep2 was considered more reliable since it was based on a more recent PhyloTree version. In study III, haplogroup information was obtained from the original publication, if available, and further confirmed with HaploGrep2.

6.1.2 Sequence alignment

Mitochondrial sequences were aligned with Sequencher v 4.10.1 (GeneCodes Inc., Ann Arbor, MI, U.S.A) (Study I), with Muscle v 3.8.31 (Edgar 2004) (Study II) and with MAFFT v7 (Katoh and Standley 2013) (Study III).

6.1.3 Basic diversity indices

For the HVR1+2 data (study I) basic diversity indices, such as number of haplotypes (A) and haplotype diversity ($\hat{H}$), were estimated with ARLEQUIN v 3.5.1.3 (Excoffier and Lischer 2010). Basic diversity indices were determined for the whole of the HVR1+2 data (N=832) and additionally for the geographical regions (North-East vs. South-West) and for the thirteen subpopulations (**Figure 9**). Differentiations between subpopulations were estimated by calculating F_{ST} and Φ_{ST} indices with 10,000 randomization steps.

6.1.4 Geographical differences in the haplogroup composition

To further evaluate the spatial variation in mitochondrial lineages, the geographical distribution of the haplogroups within Finland was estimated based on the HVR1+2 data, for which the geographical origin information was available. To assess differences between the hunter-gatherer and farmer related lineages, the HVR1+2 data was further divided into two clusters according to (Pinhasi *et al.* 2012): 1) hunter-gatherer haplogroups (HUNT; U and V), and 2) farmer haplogroups (FARM; H, J, K and T). To test the patterns of geographical differences, a linear regression analysis was performed as described in study I.

6.1.5 Identification of mitochondrial haplogroups characteristic for Finns

To estimate if Finns differ from other European populations in terms of mtDNA, complete Finnish mtDNA sequences (N=843) were compared with the GenBank database with BLAST searches (Basic Local Alignment Search Tool) (Altschul *et al.* 1990). At the time of the study, GenBank included >30,000 complete human mitochondrial sequences from all around the world. For the BLAST searches, only the haplogroups frequent enough in the Finnish dataset (N=843), were considered. Haplogroups containing less than four samples were discarded from the subsequent BLAST searches. In addition, those lineages which were known to be prevalent in other populations, such as H2a1 and V7a1, were not used for BLAST analyses. All the remaining lineages were considered to be potential Finn-characteristic haplogroups and separate BLAST searches were conducted by using one sequence from each haplogroup as a query sequence and by setting the maximum number of the target sequences to 500. For each of these target sequences, the haplogroup was assessed with HaploGrep2 and the ethnic origin for each matching sample was inferred from GenBank and/or from original publication. Those haplogroups for which Finns constituted more than 75% of all samples were considered as 'Finn-characteristic'.

6.1.6 Neighbor-Joining trees for Finn-characteristic haplogroups

In study II, Neighbor-Joining trees were reconstructed with MEGA 6 (Tamura *et al.* 2013) to display the diversity among the Finn-characteristic haplogroups. For this purpose, the most suitable substitution model was determined with PartitionFinder (Lanfear *et al.* 2012). The resulting model was TN93+Γ with shape parameter alpha value α= 0.02. These neighbor-Joining trees were rooted with a sample belonging to the haplogroup L3a (GenBank ID: JN655813) and the support values were assessed with the bootstrap method with 500 resamples.

6.1.7 Bayesian phylogenetic analyses

The focus of this thesis was to produce divergence estimates and molecular rates for the mitochondrial lineages as well as to estimate the past population sizes for the mtDNA data. Hence, the software package BEAST (Bayesian evolutionary analysis by sampling trees) based on Bayesian inference was utilized. Different versions of BEAST were used in the successive studies: v1.7.4 (study I), v1.8.2 (study II) and v2.6.2 (study III) (Drummond and Rambaut 2007; Drummond *et al.* 2012; Bouckaert *et al.* 2014; Bouckaert *et al.* 2019). BEAST produces rooted time trees and since it allows calibration of the phylogenetic tree with external information, it yields absolute calendar time estimates for branch divergence events.

All the Bayesian phylogenetic analyses followed the same procedure: the input files for BEAST were generated with BEAUTi and the actual sampling was performed with BEAST (MCMC chain length defined in each subsection). For each analysis, three to four parallel runs were performed. The consistency between the independent runs was checked with Tracer v1.6 or v1.7.1 (Rambaut *et al.* 2014; Rambaut *et al.* 2018) and runs were combined with LogCombiner, part of the BEAST software package. Effective sample size values (ESS >200) for each parameter included in the model were then inspected with Tracer v1.6 or v1.7.1 and assessed for adequacy. The phylogenetic trees were further visualized with FigTree versions v1.3.1 and v1.4.1 (http://tree.bio.ed.ac.uk/software/figtree/) and Bayesian skyline plots were visualized with Tracer.

6.1.7.1 Divergence estimates for Finn-characteristic haplogroups

To determine the divergence estimates for the Finn-characteristic mitochondrial haplogroups identified as described above, the sequence data was further divided into seven subdivisions, as the different parts of the mitochondrial genome evolve with distinctive rates. These seven categories were:

1) Nucleotides 1–576 (including HVR2)
2) Nucleotides 16,024–16,569 (including HVR1)
3) Protein coding positions at 1st codon (PC1)
4) Protein coding positions at 2nd codon (PC2)
5) Protein coding positions at 3rd codon (PC3)
6) Transfer RNAs (tRNAs)
7) Ribosomal RNAs (rRNAs)

To establish the best partition scheme and the estimation of the most suitable substitution model for each scheme, PartitionFinder v1.1.1 was used. The best partitioning for the data was '1+2', '3', '4+5' and '6+7'. The resulting substitution model was TN93+pInv+Γ for all the other schemes except '3', for which model HKY+pInv+Γ was most applicable. For each partition, the prior distribution for the mutation rate was set according to (Rieux *et al.* 2014). For a more detailed description of substitution models, see section 1.5.1 'Priors on the DNA substitution model'.

To estimate the absolute calendar years of the branching events, 12 previously published ancient mitochondrial sequences from Europe with ^{14}C-dates were used as tip calibration. These sequences belong to haplogroups U, H, T and B and cover a time span of ~31,200–700 years before present (for details, see study II supplementary table S3). A previous study has shown that approximately six ^{14}C-dated ancient DNA sequences with wide temporal distribution are sufficient to produce reliable time estimates (Molak *et al.* 2013). In addition, the uncertainty in the ^{14}C dating was taken into account,

although the effect of the sample-dating error has shown to have only minimal impact on the divergence estimates (Molak *et al.* 2013).

The phylogenetic analyses were performed with BEAST v1.8.2. Four independent runs were performed and every MCMC chain was run for 40,000,000 steps. Every 4,000th step was sampled, resulting in 10,000 trees. The first 10% of the trees were discarded as burn-in. The tree topology was assumed to be shared between all schemes and the data was further fit to three different demographic models: constant population size, exponential population growth, and the Bayesian skyline plot model. In addition, two different molecular clock models were tested (strict molecular clock and lognormal relaxed clock). To discover the best fitting demographic and clock models, Bayes factors (BF) were calculated for each model from the marginal likelihoods. Comparison of BF was performed in Tracer v1.6 and the best-fitting model was determined based on the guidelines provided in (Kass and Raftery 1995). The resulting demographic model was constant population size (Log_{10} BF 1.92 and 2.22 compared to the Bayesian skyline plot model and exponential population growth model, respectively). Strict molecular clock was considered better than the lognormal relaxed clock model (Log_{10} BF 4.05). As described in section 6.1.7, independent runs were inspected with Tracer and combined with LogCombiner. Bayesian skyline plots were visualized with Tracer.

6.1.7.2 Effective population sizes for the Finnish mitochondrial data

All effective population size estimates were determined based on the coding region (nucleotides 577–16,023) only. The control region was omitted due to the high mutation rate and possible homoplasic changes. In all analyses, the coalescent based Bayesian skyline plot was assumed as a tree prior. In study I, Bayesian skyline plots were determined for Finnish sequences belonging to the hunter-gatherer associated haplogroup U (N=86) and for the sequences belonging to the farmer associated lineage H (N=94) to compare the demographic past of Finns and other Europeans. The reference BSPs, reconstructed based on the European sequences, were obtained from (Fu *et al.* 2012). To evaluate the most suitable substitution model for the coding region, six models were fitted to the data: 1) HKY+pInv; 2) HKY+Γ; 3) HKY+pInv+Γ; 4) GTR+pInv; 5) GTR+Γ; 6) GTR+pInv+Γ. In addition, strict and lognormal relaxed molecular clocks were tested. According to the BF comparison, the strongest support was obtained for GTR with invariant sites with the relaxed clock model (see supplementary tables 4 and 5 in Översti 2014). To be able to compare the results reliably with the N_e estimates obtained for Europeans in Fu *et al.* 2012, the rate was set to 1.69×10^{-8} substitution/site/year. BSPs were constructed with BEAST v1.7.4 and for both haplogroups three independent runs were performed. MCMC chain lengths were set to 40,000,000 and 60,000,000 steps for haplogroups U and H respectively, and the first 10% of the logged steps were discarded as burn-in. Correspondingly to the previous

section, independent runs were inspected with Tracer and combined with LogCombiner. Bayesian skyline plots were visualized with Tracer.

In study II, effective population sizes were determined for two independent groups: for the sequences belonging to the Finn-characteristic mitochondrial lineages (N=281) and for the sequences belonging to the remaining haplogroups (N=562). The best-fit substitution model for the coding region was determined with PartitionFinder v1.1.1. The resulting model was TN93+pInv+Γ with four gamma categories and divided into the codon positions 1+2 and 3. The prior distribution for the mutation rate was assessed based on six previous rate estimates (Ingman *et al.* 2000; Tang *et al.* 2002; Mishmar *et al.* 2003; Ho and Endicott 2008; Soares *et al.* 2009; Fu *et al.* 2013). The resulting estimate was $\mu=1.546 \times 10^{-8}$ substitutions/site/year with standard deviation SD= $\pm$ 3.675×10^{-9} substitutions/site/year. All other settings, such as MCMC length and percentage of burn-in, and all subsequent analyses were performed as in section 6.1.7.1.

6.1.7.3 Molecular rates for mitochondrial haplogroups

The substitution model was determined for each dataset with bModelTest (Bouckaert and Drummond 2017) and all subsequent analyses were performed with BEAST v2.6.1 (Bouckaert *et al.* 2019). For each analysis, the complete mitochondrial genomes were considered as one partition. The lognormal relaxed clock model was used, as it accounts for the molecular rate variation among OTUs. A coalescent based Bayesian skyline plot was assumed as a tree prior, since it is non-parametric and hence does not require any prior knowledge of the population's demographic past. The phylogenetic tree was calibrated with ^{14}C-dates of the ancient samples. Sampling was set to 15,000,000–30,000,000 steps, depending on the sample size. All further analyses were done as described in previous parts.

For the subhaplogroups of U (U2, U4, U5a and U5b), three separate analyses including different sets of samples were carried out (see section 5.1. and also **Figure 10**):

A) Only aDNA sequences
B) aDNA sequences and contemporary sequences
C) aDNA sequences, contemporary sequences and R-root

6.1.7.4 Testing for selection

Since the differences in the selective pressure might introduce variation in the molecular rates for different collections of sequences, the codon-based Z-test of neutrality was performed for all subsets of haplogroup U. The method used to estimate the relative abundance of nonsynonymous (d_N) and synonymous (d_S) mutations was the Pamilo-Bianchi-Li method (Pamilo & Bianchi 1993; Li

1993) (Kimura 2-parameter). The variance of nonsynonymous and synonymous mutations was determined by 500 bootstrap replicates. Positions containing gaps and/or missing data were eliminated from the analysis (pairwise deletion). For each dataset, the two-tailed test was first performed to evaluate the rejection of strict neutrality (null hypothesis, $d_N = d_S$). For those datasets for which the null hypothesis was rejected ($p < 0.05$), the one-tailed test was further conducted. The aim of the one-tailed tests was to reject null hypotheses of positive and negative selection ($d_N > d_S$ and $d_N < d_S$, respectively). All the tests of neutrality were conducted in MEGA7 v7.0.26 (Kumar *et al.* 2016).

6.2 Y-chromosome

6.2.1 Haplogroup determination

Haplogroup designation for the Y-chromosomal data was performed by following the International Society of Genetic Genealogy (https://isogg.org/tree/) nomenclature (published in 2015). For the in-house dataset containing 330 samples, haplogroups were designated with two steps. Haplogroup Predictor (http://hprg.com/hapest5/) was used to preassign the haplogroups based on the Y-STR information. After the preliminary classification, samples were further genotyped for haplogroup determining SNPs. For haplogroup N, the determining SNPs included M46, M178, and L550, whereas for lineage I mutations L22, L258, and L300 were used. Genotyping of the SNPs was done with Real Time PCR by using TaqMan technology. For more detailed information on the laboratory procedures used, see study I. For the FamilyTree dataset, haplogroup predictions were obtained from Family Tree website.

6.2.2 Basic diversity indices

As for the mitochondrial HVR1+2 data, basic diversity indices were calculated for the whole dataset, for the geographical regions of South-West and North-East (SW and NE respectively), and for thirteen subpopulations with ARLEQUIN v 3.5.1.3 (Excoffier and Lischer 2010). As with the mtDNA analysis, F_{ST} and Φ_{ST} indices were estimated with 1,000 randomization steps.

7 RESULTS

7.1 Mitochondrial DNA

7.1.1 Mitochondrial DNA diversity and haplogroup composition in Finland

In study I, the mitochondrial HVR1+2 data, comprising 832 samples, displayed 384 unique haplotypes. The prevalence of haplogroups was analogous to previous results, with haplogroup H being the most common (33.2%) followed by lineage U (24.3%) (**Table 4**). Other common haplogroups were T (6.1%), J (5.5%), K (5.5%), W (3.7%) and V (3.6%), with the results being in accordance with previous studies (Meinilä *et al.* 2001; Hedman *et al.* 2007). The hunter-gatherer associated haplogroup cluster (HUNT; U and V) reached the frequency of 27.9%, whereas half of the samples belonged to the cluster of farmer related lineages (50.3%, FARM; H, J, K and T). Further, the overall haplotype diversity ($\hat{H}$=0.993±0.001) was comparable to the previous estimate ($\hat{H}$=0.992±0.002) (Hedman *et al.* 2007). The Southwestern samples showed slightly higher diversity point estimates than the Northeastern ones ($\hat{H}$=0.994±0.001 and $\hat{H}$=0.990±0.001, respectively).

In study II, among the 843 complete mitochondrial sequences retrieved from database searches altogether 240 different subhaplogroups were identified (data not shown here, see study II supplementary data file). As for the HVR1+2 data, the main haplogroup frequencies were in accordance with previously published results (**Table 4**). Since the frequencies for the complete sequence data lacking the geographical information were similar to the HVR1+2 data, for which the donor's geographical origin was available, the dataset in study II could be considered as a representative sample of Finland in the consecutive analyses.

Table 4. *Main haplogroup frequencies (%) for the mitochondrial HVR1+2 data (study I) and for the complete sequence data (study II). Haplogroup frequencies from Finnilä* et al. *2001 used as a reference.*

Haplogroup/ Cluster	Study I			Study II	Finnilä *et al. 2001*
	Whole data	North-East	South-West	Whole data	Whole data
	N=832	N=443	N=389	N=843	N=79
U	24,3	29,8	18,0	22,5	27,9
V	3,6	4,0	3,1	8,7	5,1
HUNT	27,9	33,9	21,1	31,2	33,0
H	33,2	32,3	34,2	36,9	39,1
J	5,5	2,9	8,5	6,4	5,1
K	5,5	5,0	6,2	6,0	2,5
T	6,1	5,2	7,2	4,4	2,5
FARM	50,3	45,4	56,0	53,7	49,2
I	1,2	0,9	1,5	2,8	3,8
R	1,9	2,5	1,8	0,7	-
W	3,7	4,1	3,3	7,7	10,1
X	1,8	1,4	2,3	1,4	1,3
Z	0,5	0,7	0,3	1,3	2,5
Others/ undefined	12,7	11,1	13,7	0,2	-

7.1.2 Mitochondrial East-West divergence within Finland

Based on regression analysis, the highest differences in the HUNT and FARM haplogroup frequencies were observed between Southern and Western (AL, TU, HA, VA, UU, LMO) and Northern and Eastern (MI, CF, KU, KY, NC, OU, LA) subpopulations (see **Figure 9**), with the lowest p-value $P_{HUNT} * P_{FARM} = 9.85 \times 10^{-8}$. Specifically, the haplogroup U showed NE bias, mainly dominated by the high prevalence of U5b in the northeastern parts of the country. By contrast, U5a displayed weaker, but still significant, SW affinity. In general, the HUNT cluster showed clear NE dominance, whereas the FARM cluster exhibited the opposite trend, having SW bias which was especially pronounced by haplogroup J. When considering the differences in the haplogroup ages assessed between the European and Near Eastern populations (Richards *et al.* 2000), lineages for which the age was older in Europe than in Near Eastern (U and V) had a stronger NE bias. In contrast, haplogroups which had a higher age estimate in the Near East (H, J, K, T) showed substantial Southwestern bias (R^2=0.983) (**Figure 11**).

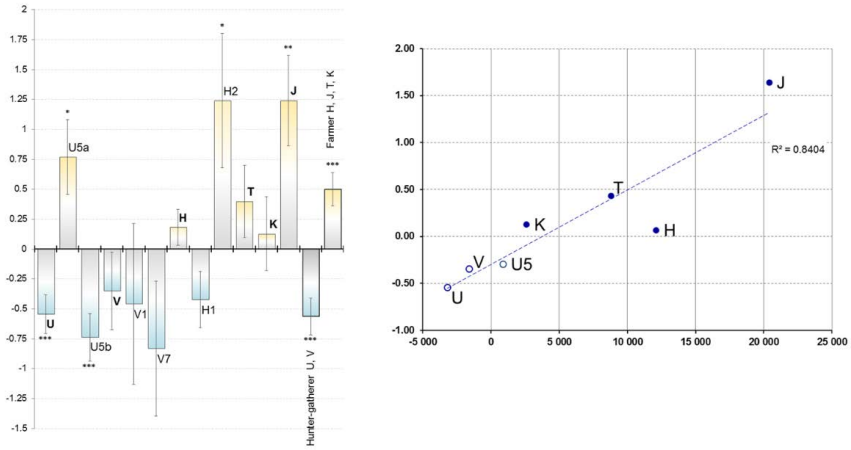

Figure 11 **a)** Geographical differences in the haplogroup frequencies based on results from the regression analysis. Above x-axis: SW dominance; below, NE dominance. Error bars represent standard deviation and asterisks denotes statistical significance. **b)** Correlation of the haplogroup ages and the observed geographical bias. On the x-axis: Haplogroup age difference in years between Near Eastern and European populations (age in Near East – age in Europe). On y-axis: geographical distribution bias in Finland (above SW bias and below NE bias). Ages obtained from (Richards *et al.* 2000). (Figure from *PLoS ONE*, CC BY 4.0 license).

7.1.3 Past population sizes for hunter-gatherer and farmer related haplogroups in Finland

To contrast the past demographic events in Finland with the signals observed for Western Europeans, Bayesian skyline plots were created for Finnish H and U sequences (N=94 and N=86 respectively). Finns showed substantially smaller effective population sizes for both haplogroups compared to the estimates obtained for Western Europeans, as expected (**Figure 12**). Furthermore, the magnitude of population expansion was considerably smaller for Finnish data. Particularly for U, the growth started later than in Europe, as late as approximately 4,000 years ago.

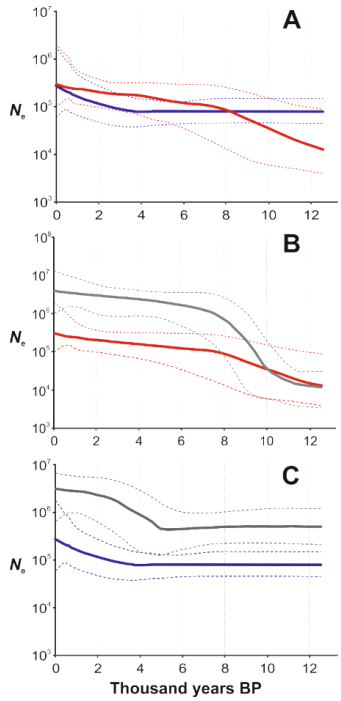

Figure 12 Bayesian skyline plots for hunter-gatherer haplogroup U and farmer haplogroup H. **a)** U (blue) and H (red) in Finland. **b)** Lineage H in Finland (red) and in Europe (grey, reference data from (Fu *et al.* 2012). **c)** Haplogroup U in Finland (blue) and in Europe (grey). In all figures, the x-axis represents time in thousand years before present and the y-axis denotes the effective female population size on a logarithmic scale. Dotted lines stand for 95% highest posterior density (HPD). (Figure from *PLoS ONE*, CC BY 4.0 license).

7.1.4 Finn-characteristic mtDNA haplogroups

When using the 75% cut-off limit, 23 mitochondrial lineages were identified as Finn-characteristic, i.e. these haplogroups were restricted to Finns and were very rare or virtually absent from other populations (**Table 5**). Out of the whole Finnish dataset comprising 843 sequences, up to 33.3% belonged to these Finn-characteristic haplogroups, which constituted both hunter-gatherer and farmer associated lineages. Since the limit of 75% was somewhat arbitrary, a more rigorous limit of 90% was also used. With this more stringent limit, 20.8% of the Finnish samples were determined as Finn-characteristic.

Table 5. *Age estimates for 23 Finn-characteristic haplogroups. Haplogroups for which the proportion of Finnish sequences exceeded 90% are bolded. HPD=highest posterior density. U5b1b2* includes samples belonging to both U5b1b2 and U5b1b2a. Similarly, V1a1a* consist of haplogroups V1a1a & V1a1a1 and W1b* of W1b & W1b1.*

Haplogroup	Median (ybp)	95% Lower HPD (ybp)	95% Upper HPD (ybp)
HUNT			
U5a2a1a	3,365	917	5,876
U5b1b1a1a	3,379	1,236	5,439
U5b1b1a1a1	1,544	310	2,395
(U5b1b2*)	5,883	2,708	8,923
U5b1b2	4,256	1,517	6,532
U5b1b2a	3,279	1,128	5,217
(V1a1a*)	4,102	1,599	6,317
V1a1a	1,942	265	3,501
V1a1a1	2,214	713	3,457
V5	4,259	1,116	7,261
V8	3,787	1,122	6,395
FARM			
H13a1a1d1	5,467	2,331	8,507
H1a2	3,488	1,039	5,974
H1f1	4,827	2,042	7,297
H1n4	3,252	861	5,560
H3h1	3,392	998	5,877
H5a1e	2,016	339	3,549
J1c2n1	1,910	472	4,490
K1c1c	5,433	2,026	8,919
Others			
I1a1a1	1,142	215	2,691
I1a1a2	973	166	2,503
I2b	1,839	331	4,722
W1a	4,676	2,070	8,057
(W1b*)	4,392	1,949	7,505
W1b	858	57	2,710
W1b1	3,688	1,782	5,913

Based on phylogenetic analyses, the divergence estimates for the Finn-characteristic haplogroups varied from ~5,900 ypb for haplogroup U5b1b2* to ~860 ybp for haplogroup W1b (**Table 5**). For the majority of the Finn-characteristic haplogroups, the median divergence estimate was approximately between 3,300 and 5,500 years before present.

To further evaluate the demographic patterns of Finn-characteristic and 'immigrant' lineages in time, BSPs were reconstructed for these two distinct

groups. The effective population size for the Finn-characteristic lineages started to decrease approximately 4,000–5,000 years before present, after a long stable phase (**Figure 13**). The lowest N_e was reached around 1,000 years ago, after which the population size started to increase rapidly, resulting in a roughly 60 times larger effective population size compared to the N_e of the lowest point. By contrast, for the non-Finn-characteristic haplogroups the N_e showed a rather constant increase between ~12,000–4,000 years before present, followed by a relatively stable population size for approximately 3,000 years. Similarto the Finn-characteristic haplogroups, immigrant lineages experienced population size expansion starting around 1,000 years, but with much more moderate growth than visible for Finn-characteristic haplogroups.

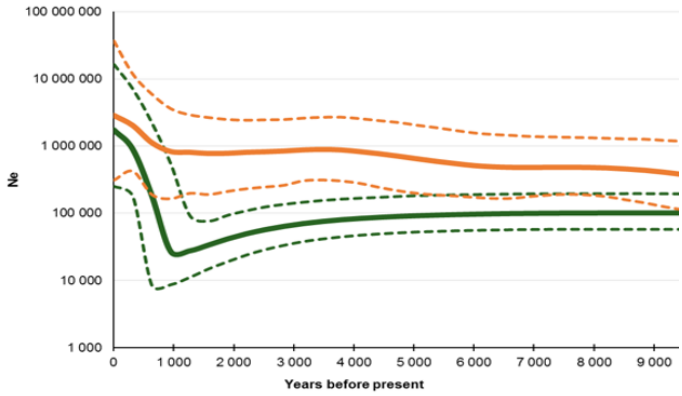

Figure 13 Effective population size estimates for the Finn-characteristic (green, N=281) and for the non-Finn-characteristic (orange, N=562) haplogroups. Continuous lines represent the mean value and dotted lines 95% HPD. The x-axis depicts years before present and the y-axis the effective population size (on logarithmic scale).

7.1.5 Variation in the mitochondrial molecular rates

To estimate possible variation in the molecular rates among mitochondrial lineages, both ancient and contemporary sequences were collected for haplogroups U2, U4, U5a and U5b. The number of samples is presented in **Table 6** and the distributions of radiocarbon dates of ancient samples are visualized in **Figure 14**.

Table 6. *Number of complete mitochondrial genomes per haplogroup used in study III.*

Haplogroup	Ancient	Contemporary
U2	19	42
U4	42	62
U5a	99	99
U5b	74	98

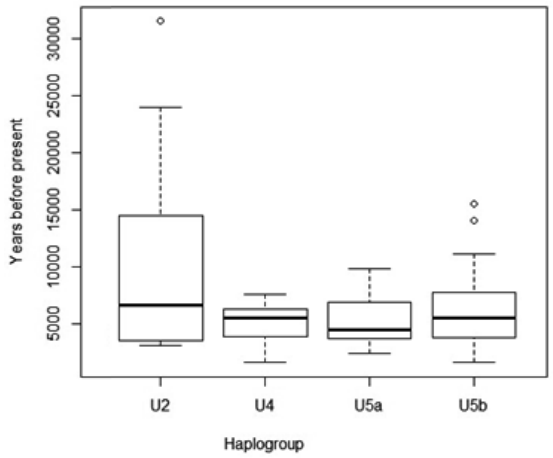

Figure 14 The temporal distributions of radiocarbon dates of the ancient samples used in the study III.

Molecular rates were determined according to three different scenarios, where scenario A) included only aDNA sequences, scenario B) contained aDNA sequences and contemporary sequences, and scenario C) where B was complemented with two ancient samples from haplogroup R, which were used as an outgroup ('R-root').

For the majority of the different datasets, the most suitable substitution model, determined with bModelTest, turned out to be HKY+Γ+pInv (see Table 2 in study III). The only exception was subhaplogroup U5b, for which the substitution model GTR+Γ+pInv gained the highest support. To test the possible bias introduced by the different substitution models, we also performed BEAST analysis for U5b with model HKY+Γ+pInv. As no differences arose (see supplementary table S3 in study III) the results can be

considered robust and not qualitatively affected by the substitution model used.

When comparing the three different scenarios (A-C) within haplogroups (**Table 7** and **Figure 15**), substantial differences arise: for each lineage, the molecular rate determined purely based on the aDNA sequences (scenario A) is considerably higher than the two other estimates. Within haplogroups, only modest differences are observed between scenarios B and C, suggesting that the addition of R-root as an outgroup has no significant impact on the molecular rate.

Table 7. *Summary statistics for the molecular rate estimates. All values are 10^{-8}. Different scenarios are A) Only aDNA sequences, B) aDNA sequences and contemporary sequences and C) aDNA sequences, contemporary sequences and R-root. 95% HPD stands for 95% highest posterior density.*

Summary	**U2**			**U4**		
statistic	A	B	C	A	B	C
Mean	5.19	4.43	4.30	5.65	3.90	3.64
Stdev of mean	1.07	0.67	0.81	1.40	0.52	0.42
Median	5.09	4.36	4.2	5.56	3.88	3.63
95% HPD	[3.27, 7.17]	[3.22, 5.78]	[3.02, 5.90]	[3.03, 8.62]	[2.88, 4.91]	[2.83, 4.47]
Summary	**U5a**			**U5b**		
statistic	A	B	C	A	B	C
Mean	5.20	3.96	3.27	3.40	2.13	2.29
Stdev of mean	0.58	0.39	0.31	0.71	0.33	0.27
Median	5.17	3.94	3.26	3.40	2.11	2.27
95% HPD	[3.58, 6.94]	[3.23, 4.74]	[2.68, 3.87]	[2.06, 4.74]	[1.53, 2.78]	[1.78, 2.83]

However, a noteworthy distinction is seen between the subhaplogroups: in scenario A) the molecular rates for U2, U4 and U5a are somewhat analogous whereas the rate for U5b is significantly lower. Furthermore, in scenarios B) and C) the rate for U5b is notably lower compared to the other lineages: U2 has evolved approximately two times faster than U5b. Haplogroups U4 and U5a show intermediate values, having around 1.4-1.8 times faster molecular rates than U5b.

To rule out the possibility that selection has introduced the observed differences in diversity within subhaplogroups, we additionally tested the possible signals of selection with Z-test (Pamilo-Bianchi-Li method, Kimura 2-parameter). The results showed that the hypothesis of selective neutrality could not be rejected (two-tailed $p > 0.149$) in the case of haplogroup U4, nor in subhaplogroups U2e, U5a and U5b, regardless of the dataset used (scenarios A, B and C) (see supplementary table S4). However, the test revealed signals of negative selection for sequences belonging to haplogroup U2 (for modern and ancient sequences, two-tailed $p = 0.019$, positive p = 1,

negative p = 0.012). As substitution rate estimates for sequences belonging to haplogroups U2 (showing negative selection), U4 (neutral) and U5a (neutral) are relatively similar, it is unlikely that selection has qualitatively affected the molecular rates obtained for sequences representing different (sub)haplogroups.

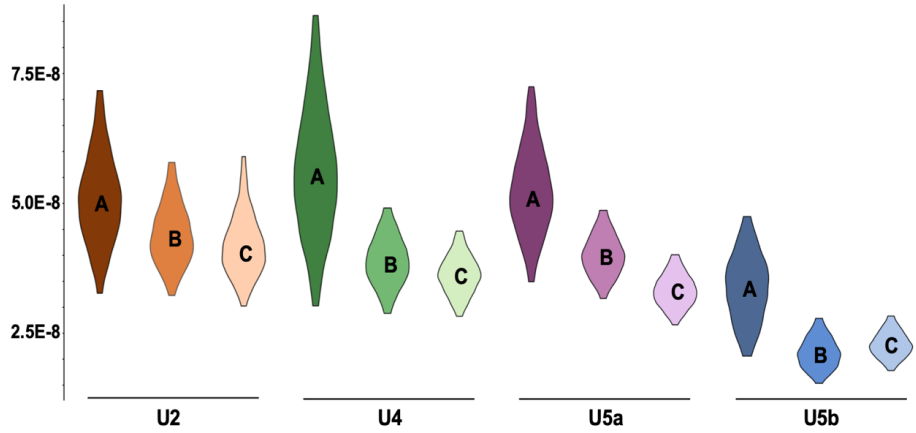

Figure 15 Molecular rates for each scenario. A) Only aDNA sequences, B) aDNA sequences and contemporary sequences and C) aDNA sequences, comtemporary sequences and R-root. Distributions represents 95% highest posterior distribution of the molecular rates (*ucld.Mean*).

7.1.6 Divergence time estimates for subhaplogroups of U

To further evaluate the possible effect of variation in the molecular rates among the mitochondrial lineages, divergence estimates were determined for sublineages of U2, U4, U5a and U5b. For comparison, divergence estimates were also collected from Soares *et al.* 2009, Malyarchuk *et al.* 2010 and Behar *et al.* 2012 (**Table 8**). In general, the divergence estimates obtained in this study are overlapping with the previous estimates, with U2 being the oldest of the lineages and U4 having the most recent divergence. For subhaplogroups U2a, U2b, U2c and U2d, the ages are determined based on relatively low sample sizes (U2a N=3, U2b N=3, U2c N=4 and U2d N=5) and hence estimates might not reflect the actual divergence times. However, the estimate for U2 (39,800 ybp [38,000, 44,600]) coincides with previous estimates.

Table 8. *Divergence estimates for sublineages of U. For comparison, estimates from Soares* et al. *2009, Malyarchuk* et al. *2010 and Behar* et al. *2012, are also presented. All values are thousands of years before present (kya). For the estimates obtained in this study, median values are presented. For mean and 95% HPD values, see supplementary table S5 in study III.*

Haplogroup	This study	Soares et al. 2009	Behar et al. 2012	Malyarchuk et al. 2010
U2	39.8	53.5	42.8	–
U2a	17.6	27.5	22.7	–
U2b	12.0	34.3	29.3	–
U2c	14.1	34.8	29.9	–
U2e	17.2	16.7	19.3	–
U2d	13.1	–	20.8	–
U4	16.9	20.9	17.5	–
U4a	13.1	–	14.9	–
U4a1	10.9	–	7.7	–
U4a2	12.0	–	8.8	–
U4b	13.4	–	12.6	–
U4b1	12.5	–	11.5	–
U4d	11.6	–	14.9	–
U5a	17.9	26.9	22.4	19.9
U5a1	15.0	18.2	16.9	16.2
U5a1a	10.3	–	12.1	–
U5a1a1	8.8	–	6.8	12.3
U5a1a2	7.6	–	10.3	–
U5a1b	7.5	–	8.4	11.2
U5a1c	13.0	–	14.6	13.0
U5a1d	10.2	–	15.1	19.0
U5a2	15.5	22.0	18.4	14.4
U5a2a	13.7	–	13.0	5.7
U5a2b	8.3	–	11.4	8.3
U5a2c	14.0	–	11.4	12.8
U5a2d	13.3	–	16.9	–
U5b	27.8	27.4	22.8	23.8
U5b1	20.7	24.0	15.5	17.7
U5b1b	17.5	–	10.8	–
U5b1c	14.1	–	10.4	12.8
U5b1d	15.1	–	11.7	–
U5b2	24.5	22.4	20.0	23.7
U5b2a	19.9	–	14.9	19.9
U5b2b	21.3	–	14.7	19.0
U5b2c	16.2	–	12.7	–
U5b3	16.2	4.3	10.5	10.6

Whereas in Soares *et al.* 2009, Malyarchuk *et al.* 2010 and Behar *et al.* 2012 for U5a and U5b the divergences estimates within the study are rather uniform, in this study an outstanding difference arises: U5b being approximately 10,000 years older than U5a. The 95% highest posterior densities for age estimates for U5a ([13,700; 21,700] ybp) and U5b ([21,600; 31,700] ybp) are barely overlapping, implying that the observed difference is significant. Subsequently, for multiple subhaplogroups of U5b the estimates presented in this study are higher than estimates in Malyarchuk *et al.* 2010 and Behar *et al.* 2012.

7.2 Y-chromosomal diversity and haplogroup composition in Finland

Among the 548 males from Finland analysed in study I, 294 unique haplotypes were retrieved, yielding an overall haplotype diversity of $\hat{H}$=0.9863±0.0019. Comparable to earlier studies (Lappalainen *et al.* 2006), lineages N-M46 (N1c1) and I-M253 (I1) were most frequent, covering altogether over 90% of the data, with haplogroup N1c1 being the most common in the NE, whereas I1 reached its highest frequencies in SW (**Table 9**). Considerable differences in haplogroup frequencies within the country are also visible at the allelic level, where the differentiation between NE and SW is significant (Φ_{ST}=0.107). Nevertheless, these two regions did not show clear differences, neither at the regional haplotype diversity (NE: $\hat{H}$=0.973±0.006 and SW: $\hat{H}$=0.987±0.002) nor at the haplotypic level (F_{ST}=0.0101).

Table 9. *Comparison of the haplogroup frequencies for Y-chromosomal data (%) (Lappalainen* et al. *2006 vs. study I).* [a] *I1 includes also I1a, I1b & I1c in Lappalainen* et al. *2016.* [b] *R1a includes also R1a1 Lappalainen* et al. *2006.* [c] *In study I, out of 584 samples for 35 individuals the sampling location could not be determined and hence these samples are included only in the 'total' dataset.*

Study	Origin, sample size	I-M253 (I1[a])	N-M46 (N1c1)	R-M420 (R1a[b])	R-M343 (R1b)	Others
Lappalainen *et al.* 2006	East, N=306	19.6	70.9	5.9	2.6	1.0
	West, N=230	41.3	41.3	8.7	5.2	3.5
	Total, N=536	28.9	58.2	7.1	3.7	2.1
Study I	NE, N=243	23.9	66.7	3.7	3.3	2.4
	SW, N=306	55.6	37.6	3.6	2.0	1.2
	Total, N=584[c]	41.4	49.5	3.8	2.7	2.6

8 DISCUSSION

This thesis is focused on the uniparental genetic markers, particularly on the maternally inherited mitochondrial DNA, which have been the target of population genetic and forensic research for several decades. Since the mitochondrial DNA does not undergo recombination causing genomic rearrangements, past coalescent events of the taxa could be directly traced. All three studies in this thesis are exploiting this feature in a Bayesian framework, making it feasible to draw conclusions on the demographic past of female populations.

Studies I and II represent the mitochondrial composition of Finland in a new light: traditional views of the Finnish mtDNA pool resembling other European populations and homogeneous distribution of haplogroups within the country are challenged. Study III highlights the variation of molecular rates among mitochondrial lineages – a question that should be explored more in detail since molecular rates can have a huge impact on key aspects of population genetics, such as divergence time and effective population size estimates.

8.1 Finns differ from other Europeans also in terms of mtDNA

The mitochondrial genome pool in Europe has shown to be relatively homogenous – the majority of lineages present among contemporary Europeans are derived from Mesolithic hunter-gatherers (U) and Neolithic farmers (H, J, K and T), and only slight differences in the haplogroup composition have been observed between the populations. Previous studies have stated that Finns are not an exception, instead we exhibit mainly 'European' haplogroups and the mtDNA diversity corresponds with other populations (Sajantila *et al.* 1995; Torroni *et al.* 1996; Hedman *et al.* 2007). Nevertheless, earlier studies have for the most part focused on the control region of the mtDNA genome, meaning that conclusions have been based on less than 7 % of the complete sequence. To achieve a more comprehensive picture of the mitochondrial content in Finland, we collected 843 complete mtDNA sequences from Finland and conducted extensive database comparisons, particularly utilizing the GenBank repository containing more than 30,000 complete mitochondrial sequences. The results revealed that out of the 240 subhaplogroups observed, 23 lineages were restricted to Finns and were either rare or completely absent from other populations. Surprisingly, up to one third (33.3%) of the Finnish samples analyzed belonged to these 'Finn-characteristic' haplogroups, hence implying a rather different pattern for the Finnish mitochondrial gene pool than previously deduced from the bare

HVR1+2 data. Finn-characteristic haplogroups include both hunter-gatherer (U and V) and farmer associated (H, J and K) lineages, though it is impossible to distinguish if these haplogroups are autochtonous or if they originally diverged elsewhere and subsequently vanished from other populations, remaining only among Finns.

When further evaluating the past population dynamics separately for the Finn-characteristic and 'immigrant' lineages, clear differences emerge: the silhouette of BSP for non-local lineages resembles those observed in many other European populations with a population size increase starting around 10,000 years ago, associated with the Neolithic revolution (Zheng *et al.* 2012). In contrast, Finn-characteristic lineages display rather constant population size until notable decline starting approximately 4,000 years ago and reaching the lowest point around 1,000 years ago, followed by a large-scale increase. Interestingly, the lowest N_e occurring approximately 1,000 years ago and the subsequent rapid population size increase has later also been supported by a genome-wide analysis study of Finns (Martin *et al.* 2018; Santiago *et al.* 2020). Temporally, this decrease is concurrent with the suggested Iron Age population bottleneck (~1,500–1,300 ybp) (Tallavaara *et al.* 2010; Oinonen *et al.* 2010; Tallavaara & Seppä 2011). In contrast, the suggested dramatic population size reduction starting ~4,700 ybp and reaching its lowest point approximately 4,000 before present (Sajantila *et al.* 1996; Sundell *et al.* 2014; Tallavaara and Pesonen 2020), is not distinguished in our analyses.

However, since the BSP is based on the coalescence and thus assumes equal reproductive success, it is particularly sensitive to population substructure. As Finns are known to display a clear genetic East-West distinction (Kittles *et al.* 1998; Lahermo *et al.* 1999; Hannelius *et al.* 2008; Salmela *et al.* 2008), the effect of possible within-country deviation was additionally tested to rule out the possibility of substructure causing the decline apparent for the Finn-characteristic lineages. Additional skyline plots were reconstructed for NE and SW Finns based on the HVR1+2 sequences from (Palo *et al.* 2009) (see study II supplementary figure S6). As no clear differences stood out in the demographic histories between these two subpopulations, the observed population size decrease for Finn-characteristic haplogroups could be considered as a real signal.

Considering the high prevalence and distinguishable demographic dynamics of Finn-characteristic lineages, one can postulate that there might have been founder effect(s) and further genetic isolation that have shaped the mitochondrial composition of Finns. Moreover, the pattern characteristic for the mitochondrial haplogroups limited to Finns is in better accordance with the observations inferred from other genetic markers. The high overall diversity observed in the Finnish mtDNA pool could be explained by the long-term genetic influx, also supported by simulation studies (Sundell *et al.* 2010; Sundell *et al.* 2014). Either way, the Finn-characteristic signal in our mitochondrial genome pool, hidden behind the overall diversity, could only be seen using fine-resolution analyses of complete mtDNA genomes. This

approach has been subsequently adopted for phylogenetic analyses of Armenians and Russians (Derenko *et al.* 2019; Malyarchuk *et al.* 2019).

8.2 Uniparental East-west distinction within Finland

Our study supported the well-established contrast in the Y-chromosomal gene pool between the Eastern and Western parts of Finland (Kittles *et al.* 1998; Lahermo *et al.* 1999; Lappalainen *et al.* 2006; Palo *et al.* 2007; Palo *et al.* 2009). In accordance with earlier studies, the most common Y-chromosomal haplogroups among Finns were N-M46 (N1c1) and I-M253 (I1), both of which demonstrated a clear geographical pattern: N-M46 reached its highest frequency in the East (67 %), whereas I-M253 was most typical in the West (56%). Based on aDNA, lineage I-M170 (I) is one of the oldest Y-chromosomal haplogroups in Europe; the earliest findings are from the Paleolithic era ~28,000 years ago from Central Europe (Fu *et al.* 2016). During the Viking Age, sublineages of I-M253 (I1) had a high occurrence rate in Scandinavia (Margaryan *et al.* 2019), and they are still prevalent among the present-day Scandinavians (Karlsson *et al.* 2006, see also **Figure 7**).

The oldest signs of N-M46 sublineages in Fennoscandia date back to 3,500 ybp, from Bolshoy Oleni Ostrov situated in the Kola Peninsula (Lamnidis *et al.* 2018). Additionally, aDNA has revealed that the sublineages of N-M46 arrived, or at least became common, in the Eastern Baltic in the late Bronze Age or early Iron Age transition, approximately 2,500 ybp (Saag *et al.* 2019). Further, this appearance of N along with the component of Siberian genetic ancestry has been temporally associated with the arrival of the proto Finnic languages to the Eastern Baltic, conceivably indicating the spread of Uralic languages from the East (Saag *et al.* 2019). Therefore, aDNA studies confirm that both Y-chromosomal haplogroups, I-M253 (I1) and N-M46 (N1c1), have been in the geographical vicinity of present-day Finland for a long time. Moreover, aDNA results support the idea of N-M46 (N1c1) expanding to Finland via an Eastern route, whereas I-DF29 (I1a) was spreading outwards from Scandinavia (Lappalainen *et al.* 2006; Palo *et al.* 2009). This Scandinavian influence, pronounced in Southwestern Finland, has further been linked with the distribution of the Corded Ware culture that was mainly present in the western parts of the country approximately 4,700–4,300 years ago (Lappalainen *et al.* 2006).

While the East-West discrepancy has been demonstrated in Y-chromosomal and autosomal markers by several independent studies, no previous analysis has been able to identify geographical differences in the mitochondrial gene pool of contemporary Finns. We showed that when clustering mitochondrial haplotypes into hunter-gatherer (HUNT: U and V) and farmer related lineages (FARM: H, J, K and T), spatial distinction becomes apparent: Southwestern Finland shows evident bias of farmer associated lineages whereas hunter-gatherer haplogroups are significantly more frequent

in Northeastern Finland than in the Southwestern parts. This deviation along the SW-NE axis follows the border of geographical differences in the Y-chromosomal haplogroup frequencies (**Figure 16**), implying that both uniparental markers bear similar within-country discrepancies, although the contrast is much less pronounced for mtDNA.

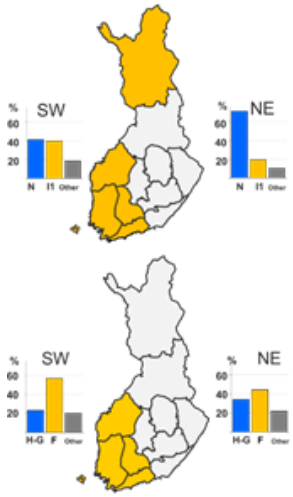

Figure 16 Within-country differences for uniparental markers in Finland. **a)** The division of Finnish provinces maximizing the Y-chromosomal haplotype differences and haplogroup frequencies for Southwestern (SW) and Northeastern (NE) Finland. **b)** The division of Finnish provinces maximizing the divergence between the mitochondrial hunter-gatherer associated lineages ('H-G': U and V) and farmer related lineages ('F': H, J, K and T). In addition, frequencies for these two clusters are presented. (Figure from *PLoS ONE*, CC BY 4.0 license).

8.3 Arrival of the agriculture-related populations to Finland

In many parts of Northern Europe, including Finland, the spread of Corded Ware Culture (CWC) has traditionally been associated with the transition to sedentary farming-based economies (see section 2 'Population history of Europe – Genetic overview'). In Finland, the distribution of CWC mainly encompassed the southwestern parts of the country approximately 4,700–4,300 ybp (Nordqvist and Häkälä 2014), implying that the culture might have arrived over the Baltic Sea from Scandinavia and/or from the Baltics. Even though some evidence of small-scale crop cultivation and animal husbandry exists from both the Neolithic and Bronze Ages in Finland (Alenius *et al.* 2013; Bläuer and Kantanen 2013; Cramp *et al.* 2014; Vanhanen *et al.* 2019), these observations are scarce and hence do not support intensive farming (see also Lahtinen *et al.* 2017). Instead, archaeobotanical studies show a noteworthy

increase in the cereal pollen records starting later, around ~100 AD (Lahtinen *et al.* 2017). The same study estimated that the cereal pollen maximum was reached during the early Medieval, as late as around 1,300 AD (Lahtinen *et al.* 2017). Correspondingly, bone findings of domesticated animals became more frequent during the middle Iron Age (Bläuer and Kantanen 2013). In addition, a distinguishable effective population size increase for sheep occurs concurrently, starting approximately 1,500 years ago (Niemi *et al.* 2018).

Universally, the introduction of animal husbandry and cereal cultivation has been connected with an increase in the human population size, although dissenting views have also been presented, stating that population sizes remained rather constant during and after the transition (see Zahid *et al.* 2016 and references therein). Nevertheless, the signal of accelerated population growth starting 10,000–8,000 ybp in Europe is evident for the mtDNA haplogroups related to Neolithic farmers (Gignoux *et al.* 2011; Fu *et al.* 2012). By contrast, for the hunter-gatherer associated lineages the population expansion during the Neolithic is less dramatic (Gignoux *et al.* 2011; Fu *et al.* 2012). The modest growth for U in Europe, starting approximately 5,000 ybp, has been interpreted to reflect the indigenous hunter-gatherer populations adopting the farming-based lifestyle, enabling a larger population size (Fu *et al.* 2012). It must be noted, however, that the Bronze Age expansion of lineages U4 and U5a might also explain, at least partially, the population size increase of U observed in Fu *et al.* 2012 (see section 3 'Population history of Europe – Mitochondrial point of view'). However, in Finnish data we observed similar tendencies as in Fu *et al.* 2012: farmer-related haplogroup H showed a gradual increase from ~9,000 onwards, whereas for U the moderate increase started later, somewhere around 4,000 years ago. Correspondingly to Western Europe, the limited growth for lineage U in the Finnish data might denote hunter-gatherer population size increase enabled by practicing small-scale animal husbandry and/or cultivation alongside hunting and fishing. However, both U and H lack the pattern of rapid expansion distinguishable in the European data, proposing that the arrival of farming related populations to Finland was less intensive and the growth of the population sizes delayed compared with many other parts of Europe. Further, BSPs for neither U nor H support a population size increase during the appearance of CWC in Finland.

Interestingly, the suggested late onset of large-scale cultivation during the late Iron Age and Medieval (Lahtinen *et al.* 2017) temporally coincide with the population size increases seen in study II. For the 'immigrant' mitochondrial lineages, this growth starts somewhere around 1,100 ybp, whereas for the Finn-characteristic haplogroups it happens slightly later, approximately ~700 years ago. The simultaneous growth in the cereal pollen records and in the mitochondrial effective population size might reflect actual population expansion enabled by more efficient land use.

The higher prevalence of farmer-related haplogroups in present-day SW Finland corresponds spatially with the distribution of CWC. It is tempting to draw a conclusion that the influx of farmer-related lineages is associated with

the CWC dispersion to Southwestern Finland and that this geographical difference in the mtDNA genome pool has remained ever since. However, a recent aDNA study showed that during the Iron Age and Medieval (~300–1,400 AD), the pattern was completely the opposite: in the Southwestern burial sites, U was the dominant mitochondrial haplogroup (58%), whereas in the Eastern sites over half of the individuals belonged to H (53%) and only 20% displayed U (Översti *et al.* 2019). Several reasons might be behind the observed contrast: assuming that haplogroups could be considered an indication of mode of subsistence, agriculturally oriented population(s) might have also arrived from the East in addition to the SW, suggesting a bidirectional spread of agriculture into Finland. On the other hand, the dispersal of CWC, and hence also the transmission of farming, might have been largely male-driven (Goldberg *et al.* 2017) leaving no traces of the migration on the mitochondrial genome pool. In this case, the high frequency of farmer-related haplogroups among contemporary SW Finns might be the result of gene flow either from Eastern Finland or over the Baltic Sea, occurring after the Iron Age. Furthermore, it is worth noting that the aDNA study (Översti *et al.* 2019) included only five burial sites (three from the South-West and two from the East) and hence might not be a representative sample of the Iron Age and Medieval maternal composition.

In conclusion, the mitochondrial DNA data supports the observations found in the pollen records (see Lahtinen *et al.* 2017). Presumably, the transition to farming occurred in Finland in several phases and was a complex, long-term process. Due to the geographical location, it is supposed that the arrival of farming populations has been less intense and migrations occurred much later than for many other European countries. The high prevalence of U among contemporary Finns implies that the contribution of hunter-gatherer related ancestry to the Finnish mitochondrial genome pool remains greater than in many other populations. Moreover, it suggests that the native hunter-gatherer populations adopted agriculture, also supported by the effective population size increase seen for Finnish U lineages. However, it is apparent that land cultivation has remained relatively small-scale until the Iron Age and Medieval, which is supported by both the pollen records and by the considerable population expansion for Finn-characteristic and non-local lineages occurring as late as ~700–1,100 years ago, respectively.

8.4 Variation among the mitochondrial molecular rates

The dating of the human mitochondrial tree has been of interest for decades and different types of external information have been used for the calibration. Traditionally, the information has been obtained from the human-chimpanzee split, geographical events and molecular rates estimated based on pedigree studies (see Endicott and Ho 2008 and references therein, see also sections 1.5.4 and 1.5.5). However, through aDNA research, tip calibration has become

more popular – radiocarbon-dates of ancient individuals can be used to infer the absolute timescale of the phylogenetic tree (for example Fu *et al.* 2013; Rieux *et al.* 2014). While several studies have focused on the comparison of the rates obtained through different calibration methods, studies evaluating the rate variation between human mtDNA lineages are scarce (Torroni *et al.* 2001; Pierron *et al.* 2011).

To assess mitochondrial lineage-specific variation more comprehensively, we constructed tip-calibrated phylogenies for subhaplogroups of U (U2, U4, U5a and U5b). For each subhaplogroup, the rate determined from the ancient sequences alone (scenario A) was higher than the two other estimates (scenario B and C). This deviation stems from the time-dependency of the molecular rates: over short time periods, all the observed mutations might not be fixed, and therefore the resulting rate seems to be higher (Ho *et al.* 2005). The same pattern has been detected from the simulation studies, where molecular rates obtained from ancient samples were accelerated compared to the substitution rates from long-term phylogenetic analyses (Ho, Kolokotronis, *et al.* 2007). Since scenarios B and C yielded somewhat similar rates, these estimates can be interpreted to reflect the substitution rates gained from the long-term phylogenetic studies.

The most interesting observation is that in all three scenarios, the rate for U5b is notably lower than for the other subhaplogroups. However, it is very unlikely that this discrepancy is caused exclusively by the differences of spontaneous mutation rates between lineages. Therefore, we propose that the distinction arises, at least to a certain degree, from different demographic histories. While subhaplogroups of U were the dominant mitochondrial lineages among the European hunter-gatherers, they were largely replaced during the Neolithic period (see section 3 'Population history of Europe – Mitochondrial point of view' and references therein). However, during the Bronze Age, U4 and U5a re-expanded along with the Yamnaya migration and whereas the prevalence of U4 and U5a increased rapidly, the frequency of U5b remained rather moderate. Accordingly, even today subhaplogroups show different distributions: U5b is common in Saami (Sajantila *et al.* 1996; Tambets *et al.* 2004), Basques (Cardoso *et al.* 2011) and Finns (Finnilä *et al.* 2001), while U5a and U4 reach their highest densities in Northeastern Europe (Malyarchuk and Derenko 2001; Bermisheva *et al.* 2002; Loogväli *et al.* 2004; Lappalainen *et al.* 2008).

Additional support for the Bronze Age expansion acting as a causal agent of the lineage-specific substitution rates is gained from the Y-chromosomal data: higher mutation rates are reported for R1b (Dupuy *et al.* 2004; Claerhout *et al.* 2018), which has also been connected with the extensive spread of Yamnaya (Haak *et al.* 2015). In fact, the (rapid) spatial expansion can involve a frequency increase of the mutations occurring on the wave front (see Peischl *et al.* 2016). Subsequently, these mutations have a higher chance of being fixed in the population than in general (for review see Excoffier *et al.* 2009; Peischl

et al. 2016). This so-called 'allele surfing' might be a plausible source of the substitution rate variation seen in particular between U5a and U5b.

Due to the discrepancy between the substitution rates for U5a and U5b, subhaplogroup U5b appears to be notably older than U5a (27,700 ybp and 17,700 ybp respectively). This contrasts with previous studies, where the divergence estimates for U5a and U5b have been relatively similar within the study (Soares *et al.* 2009; Malyarchuk *et al.* 2010; Behar *et al.* 2012). Thus, our results suggest that U5b might have emerged along with U2 and U8 before the Last Glacial Maximum, whereas U5a diverged later. Nevertheless, to date the first appearance of U5b, more ancient samples would be needed from the period predating the LGM.

9 POSSIBLE CAUSES OF ERRORS

When making inferences about past population processes based solely on the uniparental markers, a lot of information is lost which is contained in the autosomal chromosomes inherited from both parents. However, mitochondrial and Y-chromosomal DNA can provide valuable information about the female and male-specific population dynamics, and therefore, uniparentally inherited loci have preserved their solid position in population genetic studies. Nevertheless, to illustrate the full picture of the population's demographic past, all genetic markers should be analyzed in parallel.

Since several results of this thesis were obtained by assuming coalescent-based models, such as the Bayesian skyline plot, some general flaws of these methods should be raised. Effective population size (N_e) describes the size of an idealized population, which is characterized with simplified assumptions, such as random mating, equal reproductive success and non-overlapping generations (Fisher 1930; Wright 1931). It is very unlikely that a natural population fulfills all these presumptions and therefore the actual census size (N) of the population is typically much larger than the effective population size. Additionally, since the N_e estimates calculated with Bayesian skyline analysis are subject to the number of OTUs analyzed (see section 1.5.2.2. 'Bayesian skyline plot model'), the estimates of N_e should not be considered to reflect the 'real' size of the population. Instead, in studies I and II the effective population size estimates represented should be interpreted in a relative manner, as demographic trends.

In addition, a non-representative sampling has shown to bias the effective population size estimates (Kuhner 2009). In study II, the detailed geographical origin was not available for the majority of the Finnish complete mitochondrial genomes. To minimize the bias introduced by the non-random sampling, the haplogroup frequencies were compared with the frequencies obtained in study I, where the geographical details of the samples were provided. Because the frequencies were equivalent, the data collected in study II can be considered to reflect a diverse random sample of Finnish mitochondrial lineages.

Coalescent based methods are also vulnerable to population substructure, since stratification violates the assumption of panmixia. If a notable substructure exists, the derived outcome of the demographic past might be erroneous. As Finland is known to have clear spatial differences in autosomal and Y-chromosomal diversity, in study II we further assayed the possibility of significant substructure in the mitochondrial data. Therefore, BSPs were reconstructed for Southwestern and Northeastern samples based on control region sequences (data from study I). Since the NE and SW samples did not show significant differences in the past effective population sizes, it was

concluded that the population size decrease observed for Finn-characteristic lineages was not a consequence of the population substructure.

Additionally, in study II a haplogroup was defined as a 'Finn-characteristic' if more than 75% of the samples belonging to the haplogroup had a Finnish origin. As this limit was chosen somewhat arbitrarily, we also used the 90% cut-off. Both scenarios yielded the same signal: N_e for Finn-characteristic lineages experienced a distinguishable decline around 5,000–1,000 years ago, after which extensive expansion occurred (see study II supplementary figure S5). In addition, it has to be noted that not all contemporary populations are equally represented in the databases. While for the majority of European populations the number of complete mtDNA sequences deposited in publicly available repositories is overflowing, for some populations, such as certain Uralic speaker populations from the Volga-Ural area, the number of publicly available complete mitochondrial genomes is small or lacking completely.

In study III we used heterochronously sampled ^{14}C-dated ancient sequences to calibrate the molecular clock. As pointed out in the review by Ho & Shapiro (2011), the usage of heterochronous data requires somewhat even temporal sampling of the sequences and in addition a population continuity between the ancient and contemporary populations. Our data consisted of a comprehensive collection of ancient samples covering a wide timespan for each subhaplogroup (see **Figure 14**) and therefore it is unlikely that the substitution rate variation between U5a and U5b resulted from an uneven temporal distribution of the samples. Although the Bayesian inference implemented in BEAST also allows incorporation of probability distributions of tip ages, this feature was not utilized since it has been shown that taking into account the error in sample dates has only a very modest impact on the divergence time estimates (Molak *et al.* 2013; Rieux *et al.* 2014; Molak *et al.* 2015). Moreover, since the majority of the sublevel haplogroups displayed by the ancient individuals are also prevalent among present-day Europeans, maternal population continuity between the ancient and modern populations can be assumed, at least to some extent.

10 CONCLUSION

The evolutionary history of a population might involve migrations, variation in selective pressures and changes in the effective population size, among many other factors. The main interest of population genetics is to illustrate the genetic diversity of the population and identify the plausible evolutionary processes behind it. For several decades, the demographic history of humans has been reconstructed based on the genetic data obtained from modern individuals. However, when using contemporary samples, only lineages which have survived until today can be sampled, whereas extinct lineages remain concealed. Fortunately, the continuously growing field of ancient DNA research has allowed us to peek into the genetic composition of the past populations. This makes it possible to assess the relationships between past and present populations. Subsequently aDNA enables us to achieve a more comprehensive understanding of the processes that have shaped the genetic variation visible in the contemporary populations.

This thesis aims to explore the past population dynamics and demographic changes by using the maternally inherited mitochondrial DNA from modern and ancient populations. The results showed the mitochondrial genome pool of contemporary Finns in a new light: contrary to the previous studies, we were able to recognize a spatial pattern in the mitochondrial lineages within Finland and past processes which produced it. Moreover, we questioned the former view of Finnish mtDNA composition resembling other European populations and distinguished a notable proportion of mitochondrial lineages prevalent only in Finland. In addition to the mitochondrial structure of Finns, this thesis demonstrates the existence of substitution rate variation among mitochondrial haplogroups. Sublineages of U displayed a conspicuous deviation in the substitution rates, potentially arising from the different demographic histories. Given that molecular rates are essential to the majority of the conclusions drawn from the phylogenies, misspecification of the rate might consequently lead to defective interpretations of the past demographic events.

This thesis also indicates the importance of using the complete mitochondrial genomes instead of the commonly used control region. It was only this higher resolution data which enabled the detection of the mitochondrial discrepancy between Finns and the other populations. More importantly, this thesis points out the significance of approaching the data from a new angle. In study I, the within-country deviation arose only when the haplogroups were clustered into hunter-gatherer and farmer related lineages; geographical differences did not stand out when frequencies of individual haplogroups were contrasted. In study II, the division of the data into 'local' and 'non-local' lineages revealed different demographic histories for the Finn-characteristic and other haplogroups. Patterns observed in the population

sizes through time for the Finn-characteristic lineages are more in line with what is known from other genetic markers. This signal was hidden behind the overall diversity, which accounts in addition to the 'local' lineages also the variety introduced by the 'non-local' lineages. Moreover, in study III, conducting analyses separately for the subhaplogroups of U yielded lineage-characteristic substitution rates most likely reflected the differences in the past population dynamics. Overall, all the main results presented in this thesis were obtained only when the data was divided into 'non-traditional' clusters: hunter-gatherer vs. farmer haplogroups (study I), local vs. non-local lineages (study II) and individual subhaplogroup level (study III). In addition, all the mtDNA sequences employed in the three studies were collected from already published publicly available sources. Taken as a whole, this stresses that new and unconventional ideas might reveal unexpected signals of the evolutionary processes, even from the data that has already been previously analysed.

ACKNOWLEDGEMENTS

This thesis was carried out at the Faculty of Biological and Environmental Sciences at the University of Helsinki. My adventures in mitochondrial DNA started already during my master's thesis, which was conducted as a part of the Argeopop project. Throughout this project, I was introduced to the exciting past of the Finns. Subsequently, in the year 2015, the Finnish Cultural Foundation granted me with three years funding for my doctoral research, without which my project would not have been possible. During 2018-2020, my work has been enabled by the SUGRIGE project, whose main supporters are the Kone Foundation and the Jane & Aatos Erkko Foundation. In addition, I am grateful to the Finnish Konkordia Fund and the University of Helsinki for several travel grants, which provided me the opportunity to attend courses abroad and participate in numerous international conferences.

Several people have been involved in this process and I would like to express my gratitude to all of them. First of all, I would like to thank my supervisors Päivi Onkamo and Jukka Palo for their scientific direction during my master's and doctoral studies. Thank you for giving me the opportunity to carry out this interesting project. By challenging my views over the years, you have helped me to develop a deeper understanding of population genetics and of science in general. In addition, thank you for your efforts during those times when I was about to lose my "*mitovation*".

I would like to extend my thanks to my pre-examiners Agnar Helgason and Boris Malyarchuk for their constructive criticism and valuable comments on this thesis. I warmly thank Guido Barbujani for accepting my request for him to act as an opponent. Unfortunately, we were not able to share my big day in person, but hopefully we will be able to meet face to face somewhere in the near future and raise a glass of Italian red wine together! Thank you Craig Primmer for accepting the role of custos at my defence. I am also grateful to members of my thesis advisory board, Niklas Wahlberg and Jukka Corander. Thank you for all your input during my doctoral studies.

I want to thank Antti Sajantila not just for your collaboration and supervision of publications included in this thesis, but also for sharing with me your extensive knowledge of the Finnish mtDNA pool. My gratitude also goes out to all my co-authors: Anu Neuvonen, Mikko Putkonen, Tarja Sundell, Monika Stoljarova and Bruce Budowle. Without your contribution, this thesis wouldn't have been feasible.

I would like to express my gratitude to all the members of the past Argeopop project. I would especially like to thank the 005 team: Tarja, Juhana and Martin. In addition to your collaboration, I have been privileged to have your friendship for all these years. Thanks for all the moments of joy and laughter during our skumppa evenings! I thank my former colleagues in the SUGRIGE project: Elina, Jaana, Kati, Kerkko, Kerttu, AP, Nelli-Johanna,

Sanni P, Sannimari and rest of the SUGRIGE group members. It has been intriguing to work with you all. Special thanks go to Elina for several insightful discussions over the years. I would also like to express my gratitude to the past and present members of the BEDLAN project. Outi, Jenni, Kaj, Luke, Mervi, Terhi and Timo, I am thankful for your collaboration and for organizing unforgettable Seili seminars. A special credit goes to Luke, thank you for all your patience when introducing me more deeply into the Bayesian world. Particular recognition goes also to Evelyn for being a friend and for providing me the opportunity to get to know the demographic past of Peruvian populations. Minerva and Timo are warmly thanked for their cooperation and for the efficient and fun writing sessions we had. Furthermore, I wish to jointly thank all the collaborators who contributed to our 'muinaismitot' publication focusing on mitochondrial diversity among ancient Finns. I also am grateful for Kirsty for the language editing of this thesis. Soub is thanked for taking extra time to design the cover of this thesis. It turned out to be fancier than I ever imagined. In addition, members of my new tide group are warmly acknowledged for all the fascinating scientific discussions. I have already learned a lot and I am really looking forward to the future collaboration with you all.

I have been fortunate to have numerous friends supporting me along the way. Ville, Juuso and Ilari are thanked for all the unforgettable moments during all our trips when travelling with the 'high cost' travel agency. These journeys have made it possible to have moments of relaxation. Heini, thank you for all the lunch breaks in Viikki. You have given me essential encouragement. Vesa-Pekka, thank you for providing me your support and great laughs every time I needed it. I would also like to thank Ryan for being my German family. Together with Nelli, Lina, Ariane and many others, you all have helped me settle into a new environment and have been supportive during the last steps of this thesis. I also want to thank several people from all those restaurants that I have worked at for almost two decades. I have met such wonderful personalities and I would like to show special gratefulness for Mia, Sannis and Tara for being such trustworthy friends. Tiina, I am more than thankful for all the special moments we had have during our long-lasting friendship. We have shared numerous joys and sorrows; hopefully from now on there will be only joys ahead. My deepest gratitude goes to Jiihoo, who has been there for me for all these years. You have always backed me up no matter what.

Last but not least, I am grateful for my own mitochondrial lineage: thanks to my mother Carita and my brothers Tatu, Mikko, Otto, and Topi for sharing U5b1b2 with me.

Sanni

11 REFERENCES

1000_Genomes_Project_Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:56–65.

Abdellah Z, Ahmadi A, Ahmed S, Aimable M, Ainscough R, Almeida J, Almond C, Ambler A, Ambrose Karen, Ambrose Kerrie, *et al.* 2004. Finishing the euchromatic sequence of the human genome. *Nature*, 431:931–945.

Ahola M, Salo K, Mannermaa K. 2016. Almost gone: Human skeletal material from Finnish Stone Age earth graves. *Fennoscandia Archaeologica*, 33:95–122.

Alenius T, Mökkönen T, Lahelma A. 2013. Early Farming in the Northern Boreal Zone: Reassessing the History of Land Use in Southeastern Finland through High-Resolution Pollen Analysis. *Geoarchaeology* 28:1–24.

Allentoft ME, Sikora M, Sjögren KG, Rasmussen S, Rasmussen M, Stenderup J, Damgaard PB, Schroeder H, Ahlström T, Vinner L, *et al.* 2015. Population genomics of Bronze Age Eurasia. *Nature*, 522:167–172.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410.

Anderson S, Bankier AT, Barrell BG, De Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, *et al.* 1981. Sequence and organization of the human mitochondrial genome. *Nature*, 290:457–465.

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. 1999. Reanalysis and revision of the cambridge reference sequence for human mitochondrial DNA. *Nature Genetics*, 23:147.

Aris-Brosou S, Yang Z. 2002. Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. *Systematic Biology*, 51:703–714.

Balanovsky O. 2017. Toward a consensus on SNP and STR mutation rates on the human Y-chromosome. *Human Genetics*, 136:575–590.

Bachtrog D, Charlesworth B. 2001. Towards a complete sequence of the human chromosome. *Genome Biology*, 2:1–5.

Barrell BG, Bankier AT, Drouin J. 1979. A different genetic code in human mitochondria. *Nature*, 282:189–194.

Behar DM, Van Oven M, Rosset S, Metspalu M, Loogväli EL, Silva NM, Kivisild T, Torroni A, Villems R. 2012. A "copernican" reassessment of the human mitochondrial DNA tree from its root. *American Journal of Human Genetics*, 90:675–684.

Benazzi S, Slon V, Talamo S, Negrino F, Peresani M, Bailey SE, Sawyer S, Panetta D, Vicino G, Starnini E, *et al.* 2015. The makers of the Protoaurignacian and implications for Neandertal extinction. *Science*, 348:793–796.

Bermisheva MA, Tambets K, Villeins R, Khusnutdinova EK. 2002. Diversity of mitochondrial DNA haplogroups in ethnic populations of the Volga-Ural region. *Molecular Biology*, 36:802–812.

Bläuer A, Kantanen J. 2013. Transition from hunting to animal husbandry in Southern, Western and Eastern Finland: New dated osteological evidence. *Journal of Archaeological Science*, 40:1646–1666.

Boattini A, Sarno S, Mazzarisi AM, Viroli C, De Fanti S, Bini C, Larmuseau MHD, Pelotti S, Luiselli D. 2019. Estimating Y-Str Mutation Rates and Tmrca Through Deep-Rooting Italian Pedigrees. *Scientific Reports*, 9:1–12.

Bollongino R, Nehlich O, Richards MP, Orschiedt J, Thomas MG, Sell C, Fajkosŏvá Z, Powell A, Burger J. 2013. 2000 years of parallel societies in Stone Age Central Europe. *Science*, 342:479–481.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: A Software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, 10(4).

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N, *et al.* 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, 15(4).

Bouckaert RR, Drummond AJ. 2017. bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC Evolutionary Biology*, 17:1–11.

Bramanti B, Thomas MG, Haak W, Unterlaender M, Jores P, Tambets K, Antanaitis-Jacobs I, Haidle MN, Jankauskas R, Kind CJ, *et al.* 2009. Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. *Science*, 326:137–140.

Bromham L, Penny D. 2013. The modern molecular clock. *Nature Reviews Genetics*, 4:216–224.

Bromham L. 2009. Why do species vary in their rate of molecular evolution? *Biology Letters*, 5:401–404.

Bromham L, Duchêne S, Hua X, Ritchie AM, Duchêne DA, Ho SYW. 2018. Bayesian molecular dating: opening up the black box. *Biological Reviews*, 93:1165–1191.

Bronk Ramsey C. 2009. Bayesian Analysis of Radiocarbon Dates. *Radiocarbon*, 51:337–360.

Brotherton P, Haak W, Templeton J, Brandt G, Soubrier J, Jane Adler C, Richards SM, Sarkissian C Der, Ganslmeier R, Friederich S, *et al.* 2013. Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. *Nature Communications*, 4:1–11.

Brown WM, George M, Wilson AC. 1979. Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Sciences USA*, 76:1967–1971.

Browning SR, Browning BL. 2015. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *American Journal of Human Genetics*, 97:404–418.

Cardoso S, Alfonso-Sánchez MA, Valverde L, Odriozola A, Pérez-Miranda AM, Peña JA, De Pancorbo MM. 2011. The maternal legacy of Basques in northern navarre: New insights into the mitochondrial DNA diversity of the Franco-Cantabrian area. *American Journal of Physical Anthropology*, 145:480–488.

Carpelan C. 1999. Käännekohtia Suomen esihistoriassa aikavälillä 5100...1000 eKr. Pohjan poluilla; Suom. juuret nykytutkimuksen mukaan. Bidr. till

kännedom av Finlands natur och Folk 153. 153:266–274.
Chen YS, Torroni A, Excoffier L, Santachiara-Benerecetti AS, Wallace DC. 1995. Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. *American Journal of Human Genetics*, 57:133–149.
Chheda H, Palta P, Pirinen M, McCarthy S, Walter K, Koskinen S, Salomaa V, Daly M, Durbin R, Palotie A, *et al.* 2017. Whole genome view of the consequences of a population bottleneck using 2926 genome sequences from Finland and United Kingdom. *European Journal of Human Genetics*, 25:477–484.
Chowdhury B, Garai G. 2017. A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics*, 190:419–431.
Claerhout S, Vandenbosch M, Nivelle K, Gruyters L, Peeters A, Larmuseau MHD, Decorte R. 2018. Determining Y-STR mutation rates in deep-routing genealogies: Identification of haplogroup differences. *Forensic Science International: Genetics*, 34:1–10.
Clayton DA. 1992. Structure and function of the mitochondrial genome. *Journal of Inherited Metabolic Disease*, 15:439–447.
Clima R, Preste R, Calabrese C, Diroma MA, Santorsola M, Scioscia G, Simone D, Shen L, Gasparre G, Attimonelli M. 2017. HmtDB 2016: Data update, a better performing query system and human mitochondrial DNA haplogroup predictor. *Nucleic Acids Research*, 45:D698–D706.
Cramp LJE, Evershed RP, Lavento M, Halinen P, Mannermaa K, Oinonen M, Kettunen J, Perola M, Onkamo P, Heyd V. 2014. Neolithic dairy farming at the extreme of agriculture in northern Europe. *Proceedings of the Royal Society B: Biological Sciences*, 281:20140819.
Delghandi M, Utsi E, Krauss S. 1998. Saami mitochondrial DNA reveals deep maternal lineage clusters. *Human Heredity*, 48:108–114.
Derenko M, Denisova G, Malyarchuk B, Hovhannisyan A, Khachatryan Z, Hrechdakian P, Litvinov A, Yepiskoposyan L. 2019. Insights into matrilineal genetic structure, differentiation and ancestry of Armenians based on complete mitogenome data. *Molecular Genetics and Genomics*, 294:1547–1559.
Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4:699–710.
Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(1).
Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, 22:1185–1192.
Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29:1969–1973.
Dupuy BM, Olaisen B. 1996. mtDNA sequences in the Norwegian Saami and main populations. In: 16th Congress of the International Society for Forensic Haemogenetics (Internationale Gesellschaft für forensische Hämogenetik eV), Santiago de Compostela, 12–16 September 1995. Springer. p. 23–25.
Dupuy BM, Stenersen M, Egeland T, Olaisen B. 2004. Y-chromosomal microsatellite mutation rates: Differences in mutation rate between and

within loci. *Human Mutation*, 23:117–124.
Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32:1792–1797.
Edwards S, Beerli P. 2000. Perspective: Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution*, 54:1839–1854.
Ehler E, Novotńy J, Juras A, Chylénski M, Moravčik O, Pačes J. 2019. AmtDB: A database of ancient human mitochondrial genomes. *Nucleic Acids Research*, 47:D29–D32.
Ellegren H. 2000. Heterogeneous mutation processes in human microsatellite DNA sequences. *Nature Genetics*, 24:400–402.
Elson JL, Turnbull DM, Howell N. 2004. Comparative genomics and the evolution of human mitochondrial DNA: Assessing the effects of selection. *American Journal of Human Genetics*, 74:229–238.
Endicott P, Ho SYW. 2008. A Bayesian evaluation of human mitochondrial substitution rates. *American Journal of Human Genetics*, 82:895–902.
Excoffier L, Foll M, Petit RJ. 2009. Genetic consequences of range expansions. *Annual Review of Ecology, Evolution and Systematics*, 40:481–501.
Excoffier L, Lischer HEL. 2010. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, 10:564–567.
Fenner JN. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *American Journal of Physical Anthropology*, 128:415–423.
Finnilä S, Lehtonen MS, Majamaa K. 2001. Phylogenetic network for european mtDNA. *American Journal of Human Genetics*, 68:1475–1484.
Fisher R. 1930. Genetical theory of natural selection. Oxford: Clarendon Press.
Fitch WM. 1986. An estimation of the number of invariable sites is necessary for the accurate estimation of the number of nucleotide substitutions since a common ancestor. *Progress in Clinical and Biological Research*, 218:149–159.
Forster P, Harding R, Torroni A, Bandelt HJ. 1996. Origin and evolution of native American mtDNA variation: A reappraisal. *American Journal of Human Genetics*, 59:935–945.
Fort J. 2015. Demic and cultural diffusion propagated the Neolithic transition across different regions of Europe. *Journal of the Royal Society Interface*, 12:20150166.
Friedlaender J, Schurr T, Gentz F, Koki G, Friedlaender F, Horvat G, Babb P, Cerchio S, Kaestle F, Schanfield M, *et al.* 2005. Expanding Southwest Pacific mitochondrial haplogroups P and Q. *Molecular Biology and Evolution*, 22:1506–1517.
Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PLF, Aximu-Petri A, Prüfer K, De Filippo C, *et al.* 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*, 514:445–449.
Fu Q, Mittnik A, Johnson PLF, Bos K, Lari M, Bollongino R, Sun C, Giemsch L, Schmitz R, Burger J, *et al.* 2013. A revised timescale for human evolution based on ancient mitochondrial genomes. *Current Biology*, 23:553–559.
Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, Furtwängler A,

Haak W, Meyer M, Mittnik A, *et al.* 2016. The genetic history of Ice Age Europe. *Nature*, 534:200–205.
Fu Q, Rudan P, Pääbo S, Krause J. 2012. Complete mitochondrial genomes reveal neolithic expansion into Europe. *PLoS One* 7:e32473.
García-Moreno J. 2004. Is there a universal mtDNA clock for birds? *Journal of Avian Biology*, 35:465–468.
Gignoux CR, Henn BM, Mountain JL. 2011. Rapid, global demographic expansions after the origins of agriculture. *Proceedings of the National Academy of Sciences USA*, 108:6044–6049.
Goedbloed M, Vermeulen M, Fang RN, Lembring M, Wollstein A, Ballantyne K, Lao O, Brauer S, Krüger C, Roewer L, *et al.* 2009. Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpFSTR® Yfiler® PCR amplification kit. *International Journal of Legal Medicine*, 123:471–482.
Goldberg A, Günther T, Rosenberg NA, Jakobsson M. 2017. Ancient X chromosomes reveal contrasting sex bias in Neolithic and Bronze Age Eurasian migrations. *Proceedings of the National Academy of Sciences USA*, 114:2657–2662.
Graur D. 2017. An upper limit on the functional fraction of the human genome. *Genome Biology and Evolution*, 9:1880–1885.
Graves JAM. 1995. The origin and function of the mammalian Y chromosome and Y-borne genes – an evolving understanding. *BioEssays*, 17:311–320.
Günther T, Malmström H, Svensson EM, Omrak A, Sánchez-Quinto F, Kılınç GM, Krzewińska M, Eriksson G, Fraser M, Edlund H, *et al.* 2018. Population genomics of Mesolithic Scandinavia: Investigating early postglacial migration routes and high-latitude adaptation. *PLoS Biology*, 16:e2003703.
Haak W, Forster P, Bramanti B, Matsumura S, Brandt G, Tänzer M, Villems R, Renfrew C, Gronenborn D, Alt KW, *et al.* 2005. Evolution: Ancient DNA from the first European farmers in 7500-year-old neolithic sites. *Science*, 310:1016–1018.
Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, *et al.* 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, 522:207–211.
Haggren G, Halinen P, Lavento M, Raninen S, Wessman A. 2015. Muinaisuutemme jäljet: Suomen esi- ja varhaishistoria kivikaudelta keskiajalle. Gaudeamus.
Hannelius U, Salmela E, Lappalainen T, Guillot G, Lindgren CM, von Döbeln U, Lahermo P, Kere J. 2008. Population substructure in Finland and Sweden revealed by the use of spatial coordinates and a small number of unlinked autosomal SNPs. *BMC Genetics*, 9:54.
Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22:160–174.
Hastings WK. 1970. Monte carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109.
Hedman M, Brandstätter A, Pimenoff V, Sistonen P, Palo JU, Parson W, Sajantila A. 2007. Finnish mitochondrial DNA HVS-I and HVS-II population data. *Forensic Science International*, 172:171–178.

Henn BM, Gignoux CR, Feldman MW, Mountain JL. 2009. Characterizing the time dependency of human mitochondrial DNA mutation rate estimates. *Molecular Biology and Evolution*, 26:217–230.
Hervella M, Izagirre N, Alonso S, Fregel R, Alonso A, Cabrera VM, de la Rúa C. 2012. Ancient DNA from hunter-gatherer and farmer groups from Northern Spain supports a random dispersion model for the Neolithic expansion into Europe. *PLoS One*, 7:e34417.
Ho SYW, Jermiin LS. 2004. Tracing the decay of the historical signal in biological sequence data. *Systematic Biology*, 53:623–637.
Ho SYW, Phillips MJ, Cooper A, Drummond AJ. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Molecular Biology and Evolution*, 22:1561–1568.
Ho SYW, Larson G. 2006. Molecular clocks: when times are a-changin'. *TRENDS in Genetics* 22:79–83.
Ho SYW, Kolokotronis SO, Allaby RG. 2007. Elevated substitution rates estimated from ancient DNA sequences. *Biology Letters*, 3:702–705.
Ho SYW, Shapiro B, Phillips MJ, Cooper A, Drummond AJ. 2007. Evidence for time dependency of molecular rate estimates. *Systematic Biology*, 56:515–522.
Ho SYW, Endicott P. 2008. The Crucial Role of Calibration in Molecular Date Estimates for the Peopling of the Americas. *American Journal of Human Genetics,* 83:142–146.
Ho SYW, Phillips MJ. 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Systematic Biology,* 58:367–380.
Ho SYW, Lanfear R, Bromham L, Phillips MJ, Soubrier J, Rodrigo AG, Cooper A. 2011. Time-dependent rates of molecular evolution. *Molecular Ecology,* 20:3087–3101.
Ho SYW, Shapiro B. 2011. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Molecular Ecology Resources,* 11:423–434.
Ho SYW, Duchêne S, Molak M, Shapiro B. 2015. Time-dependent estimates of molecular evolutionary rates: evidence and causes. *Molecular Ecology,* 24:6007–6012.
Ho SYW. 2015. Molecular clocks. In: Rink WJ, Thompson JW, editors. Encyclopedia of scientific dating methods. Dordrecht: Springer. pp. 583–588.
Hofmanová Z, Kreutzer S, Hellenthal G, Sell C, Diekmann Y, Díez-Del-Molino D, Van Dorp L, López S, Kousathanas A, Link V, *et al.* 2016. Early farmers from across Europe directly descended from Neolithic Aegeans. *Proceedings of the National Academy of Sciences USA,* 113:6886–6891.
Howell N, Smejkal CB, Mackey DA, Chinnery PF, Turnbull DM, Herrnstadt C. 2003. The pedigree rate of sequence divergence in the human mitochondrial genome: There is a difference between phylogenetic and pedigree rates. *American Journal of Human Genetics,* 72:659–670.
Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23:254–267.
Hutchison CA, Newbold JE, Potter SS, Edgell MH. 1974. Maternal inheritance of mammalian mitochondrial DNA. *Nature,* 251:536–538.
Ilumäe AM, Reidla M, Chukhryaeva M, Järve M, Post H, Karmin M, Saag L,

Agdzhoyan A, Kushniarevich A, Litvinov S, *et al.* 2016. Human Y chromosome haplogroup N: A non-trivial time-resolved phylogeography that cuts across language families. *American Journal of Human Genetics,* 99:163–173.
Ingman M, Kaessmann H, Pääbo S, Gyllensten U. 2000. Mitochondrial genome variation and the origin of modem humans. *Nature,* 408:708–713.
Jakkula E, Rehnström K, Varilo T, Pietiläinen OPH, Paunio T, Pedersen NL, DeFaire U, Järvelin MR, Saharinen J, Freimer N, *et al.* 2008. The Genome-wide patterns of variation expose significant substructure in a founder population. *American Journal of Human Genetics,* 83:787–794.
Jobling MA, Hollox E, Hurles M, Kivisild T, Tyler-Smith C. 2014. Human evolutionary genetics. 2nd edition. New York: Garland Science, Taylor & Francis group.
Jones ER, Zariņa G, Moiseyev V, Lightfoot E, Nigst PR, Manica A, Pinhasi R, Bradley DG. 2017. The Neolithic transition in the Baltic was not driven by admixture with early European farmers. *Current Biology,* 27:576–582.
Jorde LB, Bamshad M, Rogers AR. 1998. Using mitochondrial and nuclear DNA markers to reconstruct evolution. *BioEssays* 20:126–136.
Jukes TH, Cantor CR. 1969. Evolution of protein molecules BT - Mammalian protein metabolism. *Mammalian protein metabolism.* Vol. III. p. 21–132.
Juras A, Krzewińska M, Nikitin AG, Ehler E, Chyleński M, Łukasik S, Krenz-Niedbała M, Sinika V, Piontek J, Ivanova S, *et al.* 2017. Diverse origin of mitochondrial lineages in Iron Age Black Sea Scythians. *Scientific Reports,* 7:43950.
Kaessmann H, Wiebe V, Pääbo S. 1999. Extensive nuclear DNA sequence diversity among chimpanzees. *Science,* 286:1159–1162.
Karlsson AO, Wallerström T, Götherström A, Holmlund G. 2006. Y-chromosome diversity in Sweden - A long-time perspective. *European Journal of Human Genetics,* 14:963–970.
Kass RE, Raftery AE. 1995. Bayes factors. *Journal of the American Statistical Association,* 90:773–795.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution,* 30:772–780.
Kayser M. 2017. Forensic use of Y-chromosome DNA: a general overview. *Human Genetics,* 136:621–635.
Kendall DG. 1948. On the generalized "birth-and-death" process. *Annals of Mathematical Statistics*, 19:1–15.
Kerminen S, Havulinna AS, Hellenthal G, Martin AR, Sarin AP, Perola M, Palotie A, Salomaa V, Daly MJ, Ripatti S, *et al.* 2017. Fine-scale genetic structure in Finland. *G3: Genes, Genomes, Genetics*, 7:3459–3468.
Keyser C, Bouakaze C, Crubézy E, Nikolaev VG, Montagnon D, Reis T, Ludes B. 2009. Ancient DNA provides new insights into the history of south Siberian Kurgan people. *Human Genetics*, 126:395–410.
Kimura M. 1968. Evolutionary rate at the molecular level. *Nature*, 217:624–626.
Kingman JFC. 1982. The coalescent. *Stochastic Processes and their Applications*, 13:235–248.
Kittles RA, Bergen AW, Urbanek M, Virkkunen M, Linnoila M, Goldman D,

Long JC. 1999. Autosomal, mitochondrial, and Y chromosome DNA variation in Finland: Evidence for a male-specific bottleneck. *American Journal of Physical Anthropology*, 108:381–399.
Kittles RA, Perola M, Peltonen L, Bergen AW, Aragon RA, Virkkunen M, Linnoila M, Goldman D, Long JC. 1998. Dual origins of Finns revealed by Y chromosome haplotype variation. *American Journal of Human Genetics,* 62:1171–1179.
Klug WS, Cummings MR, Spencer CA, Palladino MA, Killian D. 2016. Concepts of Genetics. 11th edition. Essex: Pearson
Krause J, Briggs AW, Kircher M, Maricic T, Zwyns N, Derevianko A, Pääbo S. 2010. A Complete mtDNA Genome of an Early Modern Human from Kostenki, Russia. *Current Biology,* 20:231–236.
Kristiansen K, Allentoft ME, Frei KM, Iversen R, Johannsen NN, Kroonen G, Pospieszny Ł, Price TD, Rasmussen S, Sjögren K-G, *et al.* 2017. Re-theorising mobility and the formation of culture and language among the Corded Ware Culture in Europe. *Antiquity*, 91:334–347.
Kuhner MK. 2009. Coalescent genealogy samplers: windows into population history. *Trends in Ecology and Evolution*, 24:86–93.
Kumar S. 2005. Molecular clocks: four decades of evolution. *Nature Reviews Genetics*, 6:654–662.
Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, 33:1870–1874.
Ladoukakis ED, Zouros E. 2001. Direct evidence for homologous recombination in mussel (Mytilus galloprovincialis) mitochondrial DNA. *Molecular Biology and Evolution,* 18:1168–1175.
Lahermo P, Sajantila A, Sistonen P, Lukka M, Aula P, Peltonen L, Savontaus ML. 1996. The genetic relationship between the Finns and the Finnish Saami (Lapps): Analysis of nuclear DNA and mtDNA. *American Journal of Human Genetics,* 58:1309–1322.
Lahermo P, Savontaus ML, Sistonen P, Béres J, De Knijff P, Aula P, Sajantila A. 1999. Y chromosomal polymorphisms reveal founding lineages in the Finns and the Saami. *European Journal of Human Genetics,* 7:447–458.
Lahtinen M, Oinonen M, Tallavaara M, Walker JWP, Rowley-Conwy P. 2017. The advance of cultivation at its northern European limit: Process or event? *Holocene* 27:427–438.
Lamnidis TC, Majander K, Jeong C, Salmela E, Wessman A, Moiseyev V, Khartanovich V, Balanovsky O, Ongyerth M, Weihmann A, *et al.* 2018. Ancient Fennoscandian genomes reveal origin and spread of Siberian ancestry in Europe. *Nature Communications*, 9:285437.
Lanfear R, Calcott B, Ho SYW, Guindon S. 2012. PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution,* 29:1695–1701.
Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D, *et al.* 2008. Correlation between Genetic and Geographic Structure in Europe. *Current Biology,* 18:1241–1248.
Lappalainen T, Koivumäki S, Salmela E, Huoponen K, Sistonen P, Savontaus ML, Lahermo P. 2006. Regional differences among the finns: A Y-chromosomal perspective. *Gene*, 376:207–215.

Lappalainen T, Laitinen V, Salmela E, Andersen P, Huoponen K, Savontaus ML, Lahermo P. 2008. Migration waves to the baltic sea region. *Annals of Human Genetics*, 72:337–348.
Lazaridis I. 2018. The evolutionary history of human populations in Europe. *Current Opinion in Genetics and Development*, 53:21–27.
Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, Fernandes D, Novak M, Gamarra B, Sirak K, *et al.* 2016. Genomic insights into the origin of farming in the ancient Near East. *Nature*, 536:419–424.
Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, *et al.* 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513:409–413.
Li WH, Sadler LA. 1991. Low nucleotide diversity in man. *Genetics*, 129:513–523.
Li W-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitutions. *Journal of Molecular Evolution*, 36:96–99.
Loogväli EL, Roostalu U, Malyarchuk BA, Derenko M V., Kivisild T, Metspalu E, Tambets K, Reidla M, Tolk HV, Parik J, *et al.* 2004. Disuniting uniformity: A pied cladistic canvas of mtDNA haplogroup H in Eurasia. *Molecular Biology and Evolution,* 21:2012–2021.
Luo S, Valencia CA, Zhang J, Lee NC, Slone J, Gui B, Wang X, Li Z, Dell S, Brown J, *et al.* 2018. Biparental inheritance of mitochondrial DNA in humans. *Proceedings of the National Academy of Sciences USA,* 115:13039–13044.
Malmström H, Gilbert MTP, Thomas MG, Brandström M, Storå J, Molnar P, Andersen PK, Bendixen C, Holmlund G, Götherström A, *et al.* 2009. Ancient DNA reveals lack of continuity between Neolithic hunter-gatherers and contemporary Scandinavians. *Current Biology,* 19:1758–1762.
Malmström H, Linderholm A, Skoglund P, Storå J, Sjödin P, Gilbert MTP, Holmlund G, Willerslev E, Jakobsson M, Lidén K, *et al.* 2015. Ancient mitochondrial DNA from the northern fringe of the Neolithic farming expansion in Europe sheds light on the dispersion process. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370:20130373.
Malyarchuk B, Derenko M, Grzybowski T, Perkova M, Rogalla U, Vanecek T, Tsybovsky I. 2010. The peopling of Europe from the mitochondrial haplogroup U5 perspective. *PLoS One* 5:e10285.
Malyarchuk BA, Derenko M V. 2001. Mitochondrial DNA variability in Russians and Ukrainians: Implication to the origin of the Eastern Slavs. *Annals of Human Genetics*, 65:63–78.
Malyarchuk BA, Litvinov AN, Derenko M V. 2019. Structure and Forming of Mitochondrial Gene Pool of Russian Population of Eastern Europe. *Russian Journal of Genetics*, 55:622–629.
Margaryan A, Lawson DJ, Sikora M, Racimo F, Rasmussen S, Moltke I, Cassidy L, Jorsboe E, Ingason A, Pedersen MW, *et al.* 2019. Population genomics of the Viking world. *Nature*, 585:390–396.
Martin AR, Karczewski KJ, Kerminen S, Kurki MI, Sarin AP, Artomov M, Eriksson JG, Esko T, Genovese G, Havulinna AS, *et al.* 2018. Haplotype sharing provides insights into fine-scale population history and disease in

Finland. *American Journal of Human Genetics,* 102:760–775.
Mathieson I, Alpaslan-Roodenberg S, Posth C, Szécsényi-Nagy A, Rohland N, Mallick S, Olalde I, Broomandkhoshbacht N, Candilio F, Cheronet O, *et al.* 2018. The genomic history of southeastern Europe. *Nature*, 555:197–203.
Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, *et al.* 2015. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, 528:499–503.
Meinilä M, Finnilä S, Majamaa K. 2001. Evidence for mtDNA admixture between the Finns and the Saami. *Human Heredity*, 52:160–170.
Merriwether DA, Clark AG, Ballinger SW, Schurr TG, Soodyall H, Jenkins T, Sherry ST, Wallace DC. 1991. The structure of human mitochondrial DNA variation. *Journal of Molecular Ecology,* 33:543–555.
Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21:1087–1092.
Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hosseini S, Brandon M, Easleyf K, Chen E, Brown MD, *et al.* 2003. Natural selection shaped regional mtDNA variation in humans. *Proceedings of the National Academy of Sciences USA,* 100:171–176.
Mittnik A, Wang CC, Pfrengle S, Daubaras M, Zariņa G, Hallgren F, Allmäe R, Khartanovich V, Moiseyev V, Tõrv M, *et al.* 2018. The genetic prehistory of the Baltic Sea region. *Nature Communications*, 9:1–11.
Moilanen J, Finnilä S, Majamaa K. 2003. Lineage-specific selection in human mtDNA: Lack of polymorphisms in a segment of *MTND5* gene in haplogroup J. *Molecular Biology and Evolution*, 20:2132–2142.
Molak M, Lorenzen ED, Shapiro B, Ho SYW. 2013. Phylogenetic estimation of timescales using ancient DNA: The effects of temporal sampling scheme and uncertainty in sample ages. *Molecular Biology and Evolution,* 30:253–262.
Molak M, Suchard MA, Ho SYW, Beilman DW, Shapiro B. 2015. Empirical calibrated radiocarbon sampler: A tool for incorporating radiocarbon-date and calibration error into Bayesian phylogenetic analyses of ancient DNA. *Molecular Ecology Resources*, 15:81–86.
Monnot S, Samuels DC, Hesters L, Frydman N, Gigarel N, Burlet P, Kerbrat V, Lamazou F, Frydman R, Benachi A, Feingold J, Rotig A, Munnich A, Bonnefont J-P, Steffan J. 2013. Mutation dependance of the mitochondrial DNA copy number in the first stages of human embryogenesis. *Human Molecular Genetics*, 22:1867–1872.
Nei M, Maruyama T, Chakraborty R. 1975. The Bottleneck Effect and Genetic Variability in Populations. *Evolution*, 29:1–11.
Nelis M, Esko T, Mägi R, Zimprich F, Toncheva D, Karachanak S, Piskáčková T, Balaščák I, Peltonen L, Jakkula E, *et al.* 2009. Genetic structure of Europeans: A view from the north-east. *PLoS One*, 4:e5472.
Nevanlinna HR. 1972. The Finnish population structure A genetic and genealogical study. *Hereditas*, 71:195–235.
Niemi M, Sajantila A, Ahola V, Vilkki J. 2018. Sheep and cattle population dynamics based on ancient and modern DNA reflects key events in the human history of the North-East Baltic Sea Region. J. Archaeol. *Scientific*

Reports, 18:169–173.
Nordqvist K, Häkälä P. 2014. Distribution of corded ware in the areas north of the Gulf of Finland - an update. *Estonian Journal of Archaeology*, 18:3–29.
Norio R. 2003. The Finnish disease heritage III: The individual diseases. *Human Genetics*, 112:470–526.
Norio R, Nevanlinna HR, Perheentupa J. 1973. Hereditary diseases in Finland; rare flora in rare soil. *Annals of Clinical Research*, 5:109–141.
Ogden TH, Rosenberg MS. 2006. Multiple sequence alignment accuracy and phylogenetic inference. *Systematic Biology*, 55:314–328.
Ohta T, Kimura M. 1971. On the constancy of the evolutionary rate of cistrons. *Journal of Molecular Evolution*, 1:18–25.
Ohta T. 1987. Very slightly deleterious mutations and the molecular clock. *Journal of Molecular Evolution*, 26:1–6.
Ohta T. 2002. Near-neutrality in evolution of genes and gene regulation. *Proceedings of the National Academy of Sciences of the United States of America*, 99:16134–16137.
Oinonen M, Pesonen P, Alenius T, Heyd V, Holmqvist-Saukkonen E, Kivimäki S, Nygrén T, Sundell T, Onkamo P. 2014. Event reconstruction through Bayesian chronology: Massive mid-Holocene lake-burst triggered large-scale ecological and cultural change. *Holocene*, 24:1419–1427.
Oinonen M, Pesonen P, Tallavaara M. 2010. Archaeological radiocarbon dates for studying the population history in eastern Fennoscandia. *Radiocarbon*, 52:393–407.
Omrak A, Günther T, Valdiosera C, Svensson EM, Malmström H, Kiesewetter H, Aylward W, Storå J, Jakobsson M, Götherström A. 2016. Genomic evidence establishes Anatolia as the source of the European Neolithic gene pool. *Current Biology,* 26:270–275.
Osada N. 2015. Genetic diversity in humans and non-human primates and its evolutionary consequences. *Genes and Genetic Systems*, 90:133–145.
Van Oven M, Van Geystelen A, Kayser M, Decorte R, Larmuseau MH. 2014. Seeing the wood for the trees: A minimal reference phylogeny for the human Y chromosome. *Human Mutation*, 35:187–191.
van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation*, 30:E386–E394.
Palo JU, Hedman M, Ulmanen I, Lukka M, Sajantila A. 2007. High degree of Y-chromosomal divergence within Finland-forensic aspects. *Forensic Science International: Genetics*, 1:120–124.
Palo JU, Ulmanen I, Lukka M, Ellonen P, Sajantila A. 2009. Genetic markers and population history: Finland revisited. *European Journal of Human Genetics,* 17:1336–1346.
Pamilo P, Bianchi NO. 1993. Evolution of the Zfx and Zfy genes: Rates and interdependence between the genes. *Molecular Biology and Evolution*, 10:271–281.
Parsons TJ, Muniec DS, Sullivan K, Woodyatt N, Alliston-Greiner R, Wilson MR, Berry DL, Holland KA, Weedn VW, Gill P, *et al.* 1997. A high observed substitution rate in the human mitochondrial DNA control region. *Nature Genetics*, 15:363–368.
Peischl S, Dupanloup I, Bosshard L, Excoffier L. 2016. Genetic surfing in

human populations: from genes to genomes. *Current Opinion in Genetics and Development*, 41:53–61.
Pennisi E. 2012. ENCODE project writes eulogy for junk DNA. *Science*, 337:1159–1161.
Pesonen P, Oinonen M, Carpelan C, Onkamo P. 2012. Early subneolithic ceramic sequences in eastern Fennoscandia - A Bayesian approach. *Radiocarbon*, 54:661–676.
Pierron D, Chang I, Arachiche A, Heiske M, Thomas O, Borlin M, Pennarun E, Murail P, Thoraval D, Rocher C, *et al.* 2011. Mutation rate switch inside Eurasian mitochondrial haplogroups: Impact of selection and consequences for dating settlement in Europe. *PLoS One*, 6:e21543.
Pikó L, Matsumoto L. 1976. Number of mitochondria and some properties of mitochondrial DNA in the mouse egg. *Developmental Biology*, 49:1–10.
Pilipenko AS, Trapezov RO, Zhuravlev AA, Molodin VI, Romaschenko AG. 2015. MtDNA haplogroup A10 lineages in Bronze Age samples suggest that ancient autochthonous human groups contributed to the specificity of the indigenous West Siberian population. *PLoS One*, 10:e0127182.
Pinhasi R, Thomas MG, Hofreiter M, Currat M, Burger J. 2012. The genetic history of Europeans. *Trends in Genetics*, 28:496–505.
Piovesan A, Antonaros F, Vitale L, Strippoli P, Pelleri MC, Caracausi M. 2019. Human protein-coding genes and gene feature statistics in 2019. *BMC Research Notes*, 12:315.
Posada D, Crandall KA. 2001. Selecting the best-fit model of nucleotide substitution. *Systematic Biology,* 50:580–601.
Posth C, Renaud G, Mittnik A, Drucker DG, Rougier H, Cupillard C, Valentin F, Thevenet C, Furtwängler A, Wißing C, *et al.* 2016. Pleistocene mitochondrial genomes suggest a single major dispersal of non-africans and a late glacial population turnover in Europe. *Current Biology,* 26:827–833.
Pybus OG, Rambaut A, Harvey PH. 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, 155:1429–1437.
Quintana-Murci L, Fellous M. 2001. The human Y chromosome: The biological role of a "functional wasteland." *Journal of Biomedicine and Biotechnology*, 2001:18–24.
Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology,* 67:901–904.
Rambaut A, Suchard MA, Xie D, Drummond AJ. 2014. Tracer v1. 6. Comput. Progr. Doc. Distrib. by author, website http//beast.bio.ed.ac.uk/Tracer
Rands CM, Meader S, Ponting CP, Lunter G. 2014. 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genetics*, 10:e1004525.
Rannala B, Cranston K. 2005. Closing gap between rocks and clocks. *Heredity*, 94:461–462.
Rannala B, Yang Z. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Ecology,* 43:304–311
Reimer PJ, Bard E, Bayliss A, Beck JW, Blackwell PG, Ramsey CB, Brown DM, Buck CE, Edwards RL, Friedrich M, *et al.* 2013. Selection and treatment

of data for radiocarbon calibration: an update to the international calibration (IntCal) criteria. *Radiocarbon*, 55:1923–1945.
Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, *et al.* 2000. Tracing european founder lineages in the near eastern mtDNA pool. *American Journal of Human Genetics,* 67:1251–1276.
Rieux A, Balloux F. 2016. Inferences from tip-calibrated phylogenies: A review and a practical guide. *Molecular Ecology,* 25:1911–1924.
Rieux A, Eriksson A, Li M, Sobkowiak B, Weinert LA, Warmuth V, Ruiz-Linares A, Manica A, Balloux F. 2014. Improved calibration of the human mitochondrial clock using ancient genomes. *Molecular Biology and Evolution,* 31:2780–2792.
Robin ED, Wong R. 1988. Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells. *Journal of Cellular Physiology*, 136:507–513.
Ruiz-Pesini E, Mishmar D, Brandon M, Procaccio V, Wallace DC. 2004. Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science*, 303:223–226.
Saag Lehti, Laneman M, Varul L, Malve M, Valk H, Razzak MA, Shirobokov IG, Khartanovich VI, Mikhaylova ER, Kushniarevich A, *et al.* 2019. The Arrival of Siberian ancestry connecting the Eastern Baltic to Uralic speakers further East. *Current Biology,* 29:1701-1711.e16.
Saag Lehti, Varul L, Scheib CL, Stenderup J, Allentoft ME, Saag Lauri, Pagani L, Reidla M, Tambets K, Metspalu E, *et al.* 2017. Extensive farming in Estonia started through a sex-biased migration from the Steppe. *Current Biology,* 27:2185-2193
Sainudiin R, Durrett RT, Aquadro CF, Nielsen R. 2004. Microsatellite mutation models: Insights from a comparison of humans and chimpanzees. *Genetics*, 168:383–395.
Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution,* 4:406–425.
Sajantila A, Lahermo P, Anttinen T, Lukka M, Sistonen P, Savontaus ML, Aula P, Beckman L, Tranebjaerg L, Gedde-Dahl T, *et al.* 1995. Genes and languages in Europe: An analysis of mitochondrial lineages. *Genome Research*, 5:42–52.
Sajantila A, Salem AH, Savolainen P, Bauer K, Gierig C, Pääbo S. 1996. Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. *Proceedings of the National Academy of Sciences USA*, 93:12035–12039.
Salmela E, Lappalainen T, Fransson I, Andersen PM, Dahlman-Wright K, Fiebig A, Sistonen P, Savontaus ML, Schreiber S, Kere J, *et al.* 2008. Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe. *PLoS One*, 3:e3519.
Sánchez-Quinto F, Schroeder H, Ramirez O, Ávila-Arcos MC, Pybus M, Olalde I, Velazquez AM V, Marcos MEP, Encinas JMV, Bertranpetit J, *et al.* 2012. Genomic affinities of two 7,000-year-old Iberian hunter-gatherers. *Current Biology,* 2:1494–1499.
Santiago E, Novo I, Pardiñas AF, Saura M, Wang J, Caballero A. 2020. Recent demographic history inferred by high-resolution analysis of linkage

disequilibrium. *Molecular Biology and Evolution*, 37:3642–3653.
Santos C, Montiel R, Sierra B, Bettencourt C, Fernandez E, Alvarez L, Lima M, Abade A, Aluja MP. 2005. Understanding differences between phylogenetic and pedigree-derived mtDNA mutation rate: A model using families from the Azores Islands (Portugal). *Molecular Biology and Evolution,* 22:1490–1505.
Sarich VM, Wilson AC. 1966. Quantitative immunochemistry and the evolution of primate albumins: Micro-complement fixation. *Science*, 154:1563–1566.
Sarich VM, Wilson AC. 1967. Immunological time scale for hominid evolution. *Science*, 158:1200–1203.
Der Sarkissian C, Balanovsky O, Brandt G, Khartanovich V, Buzhilova A, Koshel S, Zaporozhchenko V, Gronenborn D, Moiseyev V, Kolpakov E, *et al.* 2013. Ancient DNA reveals prehistoric gene-flow from Siberia in the complex human population history of North East Europe. *PLoS Genetics*, 9:e1003296.
Sarmela M. 2009. Finnish folklore atlas: Ethnic culture of Finland 2. 4th partially revised edition. Helsinki, publisher: Matti Sarmela.
Seguin-Orlando A, Korneliussen TS, Sikora M, Malaspinas AS, Manica A, Moltke I, Albrechtsen A, Ko A, Margaryan A, Moiseyev V, *et al.* 2014. Genomic structure in Europeans dating back at least 36,200 years. *Science*, 346:1113–1118.
Shi W, Ayub Q, Vermeulen M, Shao RG, Zuniga S, Van Der Gaag K, De Knijff P, Kayser M, Xue Y, Tyler-Smith C. 2010. A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations. *Molecular Biology and Evolution,* 27:385–393.
Shoemaker JS, Fitch WM. 1989. Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. *Molecular Biology and Evolution,* 6:270–289.
Sikora M, Vladimir V, Margaryan A, Damgaard PDB, Fuente C De, Renaud G, Yang MA, Fu Q, Dupanloup I, Giampoudakis K, *et al.* 2019. The population history of northeastern Siberia since the Pleistocene. *Nature*, 570:182–188.
Sjögren KG, Price TD, Kristiansen K. 2016. Diet and mobility in the corded ware of Central Europe. *PLoS One*, 11:e0155083.
Skoglund P, Malmström H, Omrak A, Raghavan M, Valdiosera C, Günther T, Hall P, Tambets K, Parik J, Sjögren KG, *et al.* 2014. Genomic diversity and admixture differs for stone-age Scandinavian foragers and farmers. *Science*, 344:747–750.
Skoglund P, Malmström H, Raghavan M, Storå J, Hall P, Willerslev E, Gilbert MTP, Götherström A, Jakobsson M. 2012. Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science*, 336:466–469.
Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, Salas A, Oppenheimer S, Macaulay V, Richards MB. 2009. Correcting for purifying selection: An improved human mitochondrial molecular clock. *American Journal of Human Genetics,* 84:740–759.
Stoneking M, Sherry ST, Redd AJ, Vigilant L. 1992. New approaches to dating suggest a recent age for the human mtDNA ancestor. *Philosophical*

Transactions of the Royal Society B: Biological Sciences, 337:167–175.
Sundell T, Heger M, Kammonen J, Onkamo P. 2010. Modelling a Neolithic population bottleneck in Finland: a genetic simulation. *Fennoscandia Archaeologica*, 27:3–19.
Sundell T, Kammonen J, Halinen P, Pesonen P, Onkamo P. 2014. Archaeology, genetics and a population bottleneck in prehistoric Finland. *Antiquity*, 88:1132–1147.
Sundell T, Kammonen J, Heger M, Palo JU, Onkamo P. 2013. Retracing prehistoric population events in Finland using simulation. In: Earl G, Sly T, Chrysanthi A, Murrieta-Flores P, Papadopoulos C, Romanowska I, Wheatley D, editors. *CAA2012: Proceedings of the 40th Conference in Computer Applications and Quantitative Methods in Archaeology*. Southampton, United Kingdom: Amsterdam University Press. p. 93–104.
Sutovsky P, Moreno RD, Ramalho-Santos J, Dominko T, Simerly C, Schatten G. 1999. Ubiquitin tag for sperm mitochondria. *Nature*, 402:371–372.
Tallavaara M, Pesonen P. 2020. Human ecodynamics in the north-west coast of Finland 10,000–2000 years ago. *Quaternary International*, 549:26–35.
Tallavaara M, Pesonen P, Oinonen M. 2010. Prehistoric population history in eastern Fennoscandia. *Journal of Archaeological Science*, 37:251–260.
Tallavaara M, Seppä H. 2012. Did the mid-Holocene environmental changes cause the boom and bust of hunter-gatherer population size in eastern Fennoscandia? *Holocene*, 22:215–225.
Tambets K, Rootsi S, Kivisild T, Help H, Serk P, Loogväli EL, Tolk HV, Reidla M, Metspalu E, Pliss L, *et al.* 2004. The western and eastern roots of the Saami - The story of genetic "outliers" told by mitochondrial DNA and Y chromosomes. *American Journal of Human Genetics,* 74:661–682.
Tambets K, Yunusbayev B, Hudjashov G, Ilumäe AM, Rootsi S, Honkola T, Vesakoski O, Atkinson Q, Skoglund P, Kushniarevich A, *et al.* 2018. Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations. *Genome Biology*, 19:1–20.
Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution,* 10:512–526.
Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution,* 28:2731–2739
Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution,* 30:2725–2729.
Tang H, Siegmund DO, Shen P, Oefner PJ, Feldman MW. 2002. Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. *Genetics*, 161:447–459.
Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86.
Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution,* 15:1647–1657.
Tong KJ, Duchêne DA, Duchêne S, Geoghegan JL, Ho SYW. 2018. A

comparison of methods for estimating substitution rates from ancient DNA sequence data. *BMC Evolutionary Biology*, 18:1–10.
Torroni A, Achilli A, Macaulay V, Richards M, Bandelt HJ. 2006. Harvesting the fruit of the human mtDNA tree. *Trends in Genetics*, 22:339–345.
Torroni A, Huoponen K, Francalacci P, Petrozzi M, Morelli L, Scozzari R, Obinu D, Savontaus ML, Wallace DC. 1996. Classification of european mtDNAs from an analysis of three European populations. *Genetics*, 144:1835–1850.
Torroni A, Rengo C, Guida V, Cruciani F, Sellitto D, Coppa A, Calderon FL, Simionati B, Valle G, Richards M, *et al.* 2001. Do the four clades of the mtDNA haplogroup L2 evolve at different rates? *American Journal of Human Genetics,* 69:1348–1356.
Vanhanen S, Gustafsson S, Ranheden H, Björck N, Kemell M, Heyd V. 2019. Maritime Hunter-Gatherers Adopt Cultivation at the Farming Extreme of Northern Europe 5000 Years Ago. *Scientific Reports,* 9:4756.
Varilo T, Laan M, Hovatta I, Wiebe V, Terwilliger JD, Peltonen L. 2000. Linkage disequilibrium in isolated populations: Finland and a young subpopulation of Kuusamo. *European Journal of Human Genetics,* 8:604–612.
Varilo T, Paunio T, Parker A, Perola M, Meyer J, Terwilliger JD, Peltonen L. 2003. The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories. *Human Molecular Genetics*, 12:51–59.
Vianello D, Sevini F, Castellani G, Lomartire L, Capri M, Franceschi C. 2013. HAPLOFIND: A new method for high-throughput mtDNA haplogroup assignment. *Human Mutation*, 34:1189–1194.
Weir JT, Schluter D. 2008. Calibrating the avian molecular clock. *Molecular Ecology*, 17:2321–2328.
Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt HJ, Kronenberg F, Salas A, Schönherr S. 2016. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Research*, 44:W58–W63.
Wilde S, Timpson A, Kirsanow K, Kaiser E, Kayser M, Unterländer M, Hollfelder N, Potekhina ID, Schier W, Thomas MG, *et al.* 2014. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proceedings of the National Academy of Sciences USA,* 111:4832–4837.
Willems T, Gymrek M, Poznik GD, Tyler-Smith C, Erlich Y. 2016. Population-scale sequencing data enable precise estimates of Y-STR mutation rates. *American Journal of Human Genetics,* 98:919–933.
Wright S. 1931. Evolution in Mendelian populations. *Genetics*, 16:97–159.
Xu Xin, Peng M, Fang Z, Xu Xiping. 2000. The direction of microsatellite mutations is dependent upon allele length. *Nature Genetics*, 24:396–399.
Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Ecology,* 39:306–314.
Yang Z, Rannala B. 2012. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, 13:303–314.
Yoder AD, Yang Z. 2000. Estimation of primate speciation dates using local molecular clocks. *Molecular Biology and Evolution,* 17:1081–1090.

Zahid HJ, Robinson E, Kelly RL. 2016. Agriculture, population growth, and statistical analysis of the radiocarbon record. *Proceedings of the National Academy of Sciences USA,* 113:931–935.

Zheng HX, Yan S, Qin ZD, Jin L. 2012. MtDNA analysis of global populations support that major population expansions began before Neolithic Time. *Scientific Reports,* 2:745.

Zuckerkandl E, Pauling LB. 1962. Molecular disease, evolution, and genetic heterogeneity. In: Kasha M, Pullman B (eds) Horizons in biochemistry. In: Kasha M, Pullman B, editors. Horizons in biochemistry. New York: Academic Press. p. 189–225.

Översti S, Majander K, Salmela E, Salo K, Arppe L, Belskiy S, Etu-Sihvola H, Laakso V, Mikkola E, Pfrengle S, *et al.* 2019. Human mitochondrial DNA lineages in Iron-Age Fennoscandia suggest incipient admixture and eastern introduction of farming-related maternal ancestry. *Scientific Reports,* 9:16883.

Översti S. 2014. Äitilinjojen fylogeneettinen hienorakenne suomalaisilla – metsästäjä-keräilijät vs. maanviljelijät (Phylogenetic fine structure of the mitochondrial lineages in Finland – Hunter-gatherers vs. farmers). Master's thesis, Department of Biology, University of Turku (language of the thesis: Finnish).